REMARKS

By the present amendment, the paragraph of the Specification appearing at page at page 26, line 26 through page 27, line 3, is amended to correct the spelling of "Altschul" and "ORF's". The paragraph of the Specification appearing at page 45, lines 1-11, is amended to capitalize the trademark "GENBANK".

Claims 5 and 9 are amended by the present amendment. Claim 5 has been amended remove "a complement of SEQ ID NO:1298". Claim 9 has been amended to delete the term "eight" as it relates to the length of the claimed probe and replace it with "forty". Support for this amendment may be found, for example, on page 10, line 29 of the Specification. Claims 10-31 are cancelled to expedite prosecution, and not to concede to the Office's rejections.

No prohibited new matter has been introduced by way of the above amendments. Applicants reserve the right to file a continuation or divisional application on any subject matter cancelled by way of this Amendment. Applicants respectfully request consideration of the subject application as amended herein.

Objections to the Specification

The Specification stands objected to by the Office because the Title of the invention is not descriptive. Specifically, the Office alleges that the Title of the invention is not descriptive since the elected claims are directed to polynucleotides. Applicants respectfully submit that Claim 1 is directed to an isolated nucleic acid sequence encoding a polypeptide of SEQ ID NO:3218. Accordingly, Claim 1 includes a polypeptide encoded by a nucleic acid sequence of Applicant's invention and the Title of the invention, which recites "Nucleic Acid and Amino Acid Sequences," is descriptive. Reconsideration and withdrawal of the objection is respectfully requested.

The Specification is also objected to because the trademark GENBANK, appearing in the Specification is not capitalized and the terms "andORF's" and "Altshal" appear to be misspelled. By way of the present amendments, Applicants have corrected such terms and respectfully request reconsideration and withdrawal of the present objections to the Specification.

Rejections under 35 U.S.C. § 101

Claims 1-10 stand rejected under 35 U.S.C. § 101 as allegedly not supported by a specific, substantial and credible utility. Applicants note that Claim 10 has been cancelled by way of the present amendment, thereby rendering any rejection of that claim moot. The rejections of Claims 1-9 are respectfully traversed.

The Manual of Patent Examining Procedure (MPEP) states at § 2107.01, that research tools can be "useful" in a patent sense:

> Many research tools such as ... nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the invention is in fact "useful" in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified substantial utility and inventions whose asserted utility requires further research to identify or reasonably confirm.

Therefore, nucleotide sequencing techniques, which can include microbial genomic databases containing nucleic acid sequences, amino acid sequences and sequence homology information of bacterial genes that are, in turn, useful in the functional analysis of the bacterial genome, can meet the utility requirement of 35 U.S.C.§ 101 if, for example, the nucleic acid sequences and proteins encoded by the nucleic acid sequences have specific, substantial and credible utility, such as in the development of antibiotics, diagnostics, vaccines and drugs to treat humans afflicted with infection caused by the bacteria.

Further, the MPEP states, at § 2107.02B, that the utility of 35 U.S.C. § 101 is met, even if a specific, substantial and credible utility for the claimed invention is not asserted in the specification, if such utility is well-established:

> An invention has a well-established utility if (i) a person of ordinary skill in the art would immediately appreciate why the invention is useful based on the characteristics of the invention (e.g., properties or applications of a product or process), and (ii) the utility is specific, substantial, and credible. If an invention has

> a well-established utility, rejections under 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph, based on lack of utility should not be imposed. (citations omitted).

In addition, the guidelines for examination of patent applications under 35 U.S.C. § 101, "utility" requirement, as shown in the Federal Register, Vol. 66, No. 4, 1092-1099, at 1097, provides that:

> Only one specific, substantial and credible utility is required to satisfy the statutory requirement. Where one or more well-established utilities would have been readily apparent to those of skill in the art at the time of the invention, an applicant may rely on any one of those utilities without prejudice. (emphasis added)

Applicants respectfully submit that one or more well-established utilities would be readily apparent to one of skill in the art. In particular, one of skill in the art would recognize and appreciate the utility of the claimed invention for the purposes of developing new drug targets, diagnostics and therapeutics. The Specification, for example at page 2, lines 9-26 provides that *M. catarrhalis* is the most important lower respiratory tract after *S. pneumoniae* and *H. influenzae* (Doren, G., *et al*, *Diagn. Microbiol. Infect. Dis.* 4:191-201 (1986)), and in some hospitals *M. catarrhalis* accounts for half of all respiratory infections (Bluesone, C., *et al*, *Pediatr. Infect. Dis. J.* 11:S7-S11 (1992)). Further complicating the treatment of infections caused by *M. catarrhalis*, as stated on page 2, line 27 through page 3, line 4 of the Specification, *M. catarrhalis* has become resistant to many antibiotics, including vancomycin. (Doern, G., *et al*, *Antimicob. Agents Chemother.* 40:2884-2886 (1996); Wallace, RJ. *Am. J. Med.* 88:46S-50S (1990); Hoppe, HL. *Am .J. Health. Syst. Pharm.* 55:1881-97 (1998)).

The utility of identifying essential genes for *M. catarrhalis*, which can be employed, for example, to develop new drug targets, diagnostics and therapeutics for the antibiotic resistant *M. catarrhalis* is also discussed, for example, on page 3, lines 12-23 of the Specification:

> The present invention fulfills the need for diagnostic tools and therapeutics by providing bacterial-specific compositions and methods for detecting *Moraxella* species including *M. catarrhalis*, as well as compositions and

> methods useful for treating and preventing *Moraxella* infection, in particular, *M. catarrhalis* infection, in vertebrates including mammals.
>
> The present invention encompasses isolated nucleic acids and polypeptides derived from *M. catarrhalis* that are useful as reagents for diagnosis of bacterial disease, components of effective antibacterial vaccines, and/or as targets for antibacterial drugs including anti-*M. catarrhalis* drugs. They can also be used to detect the presence of *M. catarrhalis* and other *Moraxella* species in a sample; and in screening compounds for the ability to interfere with the *M. catarrhalis* life cycle or to inhibit *M. catarrhalis* infection.

The usefulness of identifying genes for new therapeutics and diagnostics would be readily apparent to one of skill in the art at the time of Applicant's invention. Specifically, the utility of genome sequencing information from microbial pathogens, in particular antibiotic resistant bacterial pathogens, is well-established in the art. For example, Moir, *et al., Antimicrob. Agents Chemother.* 43: 439-446 (1999), a copy of which is attached hereto as Exhibit 1, provides that genomic sequence information has provided a wealth of information to assist in the development of strategies for antimicrobial drug discovery, particularly in antibiotic-resistant bacteria. Specifically, on page 439 Moir, *et al.* provides:

> Thus, there is little doubt that new antibiotics are needed to combat the growing problem of antibiotic-resistant bacteria, and targeting of new pathways will likely play an important role in discovery of these new antibiotics. In fact, a number of crucial cellular pathways, such as secretion, cell division and many metabolic functions remain untargeted today. In the last 3 years, high-throughput automated random genomic DNA sequencing together with robust fragment assembly tools has delivered a wealth of genomic sequence information to assist in the search for new targets. In many cases, entire biochemical pathways can be reconstructed and compared in different pathogens.

In addition, Tatusov, *et al., Science* 278: 631-637 (1997), a copy of which is attached hereto as Exhibit 2, provides on page 631, that comparisons of complete genomic sequences of bacteria are useful and can be critically important to the development of targets for new antibiotics:

> With multiple genome sequences, it is possible to delineate protein families that are highly conserved in one domain of life but are missing in the others. Such information may be critically important: For example, the families that are conserved among bacteria but are missing in eukaryotes comprise the pool of potential targets for broad-spectrum antibiotics.

Smith DR, *TIBTECH* 14: 290-293(1996), a copy of which is attached hereto as Exhibit 3, provides that microbial genome sequence information is useful in new strategies for identifying therapeutics and vaccine development. Specifically, on page 293 Smith provides:

> The techniques described in the previous section can be used to identify genes in specific functional categories that may represent good targets for drug or vaccine development. In general, when developing new antibiotics, one is interested in genes that are essential under all growth conditions (and preferably even in quiescent cells), and for which inhibitors with useful chemical properties, such as permeability and low toxicity, can be identified. One advantage of having the entire sequence of a genome is that targets can be prioritized in terms of their activities and the properties of compounds that are known to interact with them.

In addition, Smith states, on pages 291-292, that the first task in identifying new strategies for therapeutics and vaccine targets is to identify genes of the microbial organism and that the second task is identifying sequence homology which is useful in the analysis of gene products. Specifically, on page 292 Smith provides:

> The second phase in the analysis of bacterial genomes is to identify the function of as many genes as possible. Currently, sequence homology is the most powerful tool. A high degree of homology between the putative translation product of a newly identified gene and an enzyme whose function has been thoroughly studied in other organisms, provides strong support for the function of that protein.

Applicants respectfully submit that the usefulness of identifying genes for new therapeutics and diagnostics would be readily apparent to one of skill in the art at the time of Applicant's invention.

In addition to having identified the claimed nucleic acid of *M. catarrhalis*, Applicants have also provided the identified nucleic acid sequence homology, thereby providing strong support for the function of the claimed protein. Specifically, Table 2 of the Specification provides that the amino acid sequence SEQ ID NO:3218, encoded by nucleic acid sequence SEQ ID NO:1298, asserts homology with *H. influenzae* acetylglucosamine-1-phosphate uridyltransferase, an essential protein in the synthetic pathway of *H. influenzae*. Furthermore, Table 2 provides the score and probability (determined by the BLASTP2 algorithm) and homology match to *H. influenzae* acetylglucosamine-1-phosphate uridyltransferase for Applicant's claimed invention, the amino acid sequence (SEQ ID NO:3218) encoded by SEQ ID NO:1298, with the amino acid sequence of *H. influenzae* acetylglucosamine-1-phosphate uridyltransferase.

As stated in the Federal Register at Vol. 66, No. 4, at page 1096:

> More specifically, when a patent application claiming a nucleic acid asserts a specific, substantial, and credible utility, and bases the assertion upon homology to existing nucleic acids or proteins having an accepted utility, the asserted utility must be accepted by the examiner unless the Office has sufficient evidence or sound scientific reasoning to rebut such an assertion. "[A] 'rigorous correlation' need not be shown in order to establish practical utility; 'reasonable correlation' is sufficient." (citations omitted).

Thus, nucleic acid sequences and their encoded amino acid sequences, which are homologous to known sequences with accepted utility, can meet the utility requirement of 35 U.S.C. § 101, if, for example, the homologous nucleic acid and amino acid sequences have accepted utility and the nucleic acid and amino acids sequences of the invention assert a specific, substantial and credible utility, such as the function of the homologous protein.

As shown in Table 2 of the Specification, Applicants' claimed invention, amino acid sequence SEQ ID NO:3218, encoded by nucleic acid sequence SEQ ID NO:1298, asserts homology with *H. influenzae* acetylglucosamine-1-phosphate uridyltransferase, an essential protein in the synthetic pathway of *H. influenzae*.

In addition to the assertions in Table 2, assertions of utility to homologous sequences can be found in the Specification, for example, page 2, line 9 through page 3, line 9; page 39, lines 28-29; and page 44, lines 24-25. A description of Table 2 and well-known software sequence comparisons programs, which were employed to identify the homologous sequences, can also be found in the Specification, for example, on page 6, lines 24-29.

Applicants' claimed invention provides nucleic acid sequences which encode polypeptides for use in new strategies for diagnostics and therapeutics. Specifically, Applicants' claimed invention, which is part of a microbial genomic database of sequences generally referred to by Moir, *et al.,* Tatusov, *et al.,* and Smith, include a wide variety of nucleic acid sequences which encode proteins that share homology with known proteins that have utility, several of which have been shown to be essential to life of bacteria. In particular, the claimed subject matter of the instant application has homology with a protein involved in an essential synthetic pathway in *H. influenzae.*

In addition to statements made by the Applicants with respect to utility of the claimed invention, for example, page 3, lines 12-23 and page 4, lines 2-21 of the Specification, substantial utility is well-established in view of the statements of Moir, *et al.,* Tatusov, *et al.,* and Smith that microbial genomic databases, containing nucleic acid and amino acid sequences, are useful in diagnostics, the development of new vaccines and in the search for antimicrobial drug discovery. The utility is well-established and credible, under 35 U.S.C. § 101, when assessed from the perspective of one of ordinary skill in the art in view of the disclosure and the statements of Moir, *et al.,* Tatusov, *et al.,* and Smith, that microbial genomic databases, having nucleic acid and amino acid sequences, have afforded "new tools to take advantage of genomic sequence information in the drug discovery process". (Moir, *et al.,* at 439).

The Office has not provided any evidence to rebut the assertion that the claimed nucleic acid sequence would not have the requisite utility, at least as a target for drug development and development of diagnostic tools. Furthermore, the Office has not provided any evidence suggesting that one or more well-established utilities would not have been readily apparent to one of skill in the art at the time of the invention. Therefore, Applicants respectfully submit that the claimed invention meets the requirements of 35 U.S.C. § 101 and withdrawal of the rejection is respectfully requested.

Rejections under 35 U.S.C. § 112, first paragraph

The Office has rejected Claims 1-10 under 35 U.S.C. § 112, first paragraph, as allegedly "...not being supported by a specific, substantial, and credible utility...one skilled in the art clearly would not know how to use the claimed invention." The Office has also rejected Claims 8 and 10 under 35 U.S.C. § 112, first paragraph, as allegedly containing subject matter which was not described in the Specification in such a way as to reasonably convey to one of skill in the art that Applicants had possession of the claimed invention at the time of filing. Applicants note that in order to expedite prosecution, and not to concede to the Office's rejection, Applicants have cancelled Claim 10, thus rendering any rejection of that claim moot. The rejections of Claims 1-9 are respectfully traversed.

The Office should "not impose a 35 U.S.C. § 112, first paragraph, rejection grounded on a 'lack of utility' basis unless a 35 U.S.C. § 101 rejection is proper." MPEP § 2107 (IV) at 2100-36. As discussed *supra*, Claims 1-9 comply with the utility requirement set forth in 35 U.S.C. § 101. Accordingly, withdrawal of the rejection is respectfully requested.

Claim 8 and 10 are also rejected under 35 U.S.C. § 112, first paragraph, as allegedly containing subject matter not described in the specification in such a way as to reasonably convey to one of skill in the relevant art that Applicants had possession of the claimed invention. Applicants have cancelled Claim 10, thus rendering any rejection of that claim moot. Specifically, the Office alleges that Claim 8, which depends on Claim 5, is drawn to a method of producing an *M. catarrhalis* polypeptide encoded by a complement of SEQ ID NO:1298. In order to expedite prosecution, and not to concede to the Office's rejection, Applicants have amended Claim 5, to remove the reference to "a complement" of SEQ ID NO:1298, thus obviating the rejection of Claim 8. Applicants respectfully submit that the claimed invention meets the requirements of 35 U.S.C. § 112, first paragraph and withdrawal of the rejection is respectfully requested.

Rejections under 35 U.S.C. § 102(b)

Claims 9-10 stand rejected under 35 U.S.C. § 102(b) as allegedly being anticipated by Wedler, *et al.* (Database sequence GenBank accession number Z72861, version Z72861.1,

8/11/1997). Specifically, the Office alleges that Wedler, *et al.* discloses a nucleic acid comprising a sequence that has 21 contiguous nucleotides of the claimed invention. In order to expedite prosecution, and not to concede to the Office's rejection, Claim 10 has been cancelled, thus rendering any rejection of that claim moot. The rejection of Claim 9 is respectfully traversed.

Anticipation requires the disclosure in a single prior art reference of each element of the claim under consideration. *W.L. Gore & Associates v. Garlock, Inc.*, 220 USPQ 303, 313 (Fed. Cir. 1983), *cert. denied*, 469 U.S. 851 (1984); *Connell v. Sears Roebuck & Co.*, 220 USPQ 193, 198 (Fed. Cir. 1983); *Verdegaal Bros. v. Union Oil Co. of California,* 2 USPQ2d 1051, 1053 (Fed. Cir. 1987); *In re Spada*, 15 USPQ2d 1655 (Fed. Cir. 1990); MPEP § 2131.

In order to expedite prosecution, Applicants have amended Claim 9 to recite a probe comprising forty contiguous nucleotides of SEQ ID NO:1298, thus obviating the rejection of Claim 9. Applicants respectfully submit that the claimed invention meets the requirements of 35 U.S.C. § 102(b) and withdrawal of the rejection is respectfully requested.

CONCLUSION

A general authorization is granted to hereby charge any fees or deficiencies to Deposit Account No. 501040. In view of the amendments and remarks, it is believed that all claims are in condition for allowance, and it is respectfully requested that the application be passed to issue. If the Examiner feels that a telephone call would expedite the prosecution of this case, the Examiner is invited to call the undersigned at (781) 398-2548.

Respectfully submitted,
OSCIENT PHARMACEUTICALS CORPORATION

By_____
Robert L. Spadafora, Esq.
Registration No. 46,197
Telephone (781) 398-2300
Facsimile (781) 398-2530

Waltham, Massachusetts 02451
Dated: 8/6/07

ANTIMICROBIAL AGENTS AND CHEMOTHERAPY, Mar. 1999, p. 439–446
0066-4804/99/$04.00+0

Vol. 43, No. 3

MINIREVIEW

Genomics and Antimicrobial Drug Discovery

DONALD T. MOIR,[1] KAREN J. SHAW,[2] ROBERTA S. HARE,[2] AND GERALD F. VOVIS[1]*

Pathogen Genetics Department, Genome Therapeutics Corporation, Waltham, Massachusetts 02453-8443,[1] and Chemotherapy and Molecular Genetics, Schering-Plough Research Institute, Kenilworth, New Jersey 07033-0539[2]

INTRODUCTION

The increasing frequency of nosocomial infections due to methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus faecium* (VRE) and the fear that high-level vancomycin resistance will eventually spread to staphylococci underscore the need for vigilance in the continuing war against pathogenic microbes (18, 39). Current widely used antibiotics are targeted at a surprisingly small number of vital cellular functions: cell wall, DNA, RNA, and protein biosynthesis (Table 1), and instances of resistance to these antibiotics are widespread and well documented (48). Thus, there is little doubt that new antibiotics are needed to combat the growing problem of antibiotic-resistant bacteria, and targeting of new pathways will likely play an important role in discovery of these new antibiotics. In fact, a number of crucial cellular pathways, such as secretion, cell division, and many metabolic functions, remain untargeted today. In the last 3 years, high-throughput automated random genomic DNA sequencing together with robust fragment assembly tools has delivered a wealth of genomic sequence information to assist in the search for new targets. In many cases, entire biochemical pathways can be reconstructed and compared in different pathogens. The purpose of this minireview is to indicate where this information can be found, to outline some of the ways in which it can be used, and to describe new tools to take advantage of genomic sequence information in the drug discovery process.

Each potential new antibiotic must meet a number of criteria before it is approved for use, and the choice of an appropriate target is the first step in this process. It is helpful to review the utility of genomic information with regard to some of the key criteria which antimicrobial targets must meet. In general, (i) a target should provide adequate selectivity and spectrum, yielding a drug which is specific or highly selective against the microbe with respect to the human host but also active against the desired spectrum of pathogens; (ii) a target should be essential for growth or viability of the pathogen, at least essential under conditions of infection; and (iii) something about the function of the target should be known so that assays and high-throughput screens can be built. Identification of potential new targets can proceed from any one of these criteria, but ultimately all must be met by a successful target. For example, a variety of methods may be used to find genes which are essential for the survival of an organism under defined conditions or which are necessary for infectivity in an animal model. Comparative genomics may be used to identify potential targets which are shared across multiple microbial species. Several tools, primarily sequence similarity based, may be used to predict the function of most genes so that specific pathways can be targeted. As discussed below, genomic sequence information provides assistance in all of these areas: selectivity, spectrum, functionality, and essentiality (Fig. 1).

CURRENT RESOURCES FOR GENOMIC SEQUENCE AND FUNCTIONALITY INFORMATION

Numerous databases are now available which contain both sequence and functionality information. Most of these are accessible over the Internet through convenient Web browser interfaces. Many also permit downloading of sequence information for use on local servers. Sequence databases now contain the nucleotide and predicted amino acid sequences of virtually every gene in the model microbes *Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae* as well as in a variety of other bacteria (Table 2; a version of this table is updated regularly by The Institute for Genomic Research [TIGR] on their Web site: http://www.tigr.org/tdb/mdb/mdb.html). These databases are the result of extensive analysis of the genomic sequences of those organisms. Open reading frames have been analyzed by sequence comparison and by codon usage to identify those which are most likely to represent transcribed genes. Putative functions have been assigned to slightly more than half of the genes in the model organisms based on sequence comparisons to genes of known function in other organisms, shared sequence motifs, or clustering of sequences into related families. Databases such as EcoCyc, KEGG, and WIT present these data in an organized and useful manner (see Table 3).

Recently, some commercial databases have also become available for nonexclusive use by commercial subscribers. These databases generally also provide sequence information not available in public databases and comparative software and analysis tools for convenient analysis of the data. For example, the results of prerun sequence similarity searches may be stored to provide rapid answers to complex comparative genomic queries by a subscriber. Finally, several Web-accessible sites offer useful tools for sequence analysis via sequence similarity searches, motif searches, and structural comparisons. Examples of relevant Internet sites providing databases of sequence and functionality information and research tools are described in Table 3.

The next advance in microbial genomics will be the availability of the complete genomic sequence from multiple strains of a single bacterial pathogen. The discovery of genes conserved in multiple pathogenic strains or the recognition of genes found only in the most virulent strains are examples of the power such genomic comparisons will provide. Sequence for a second strain of *Helicobacter pylori* has appeared and

* Corresponding author. Mailing address: Genome Therapeutics Corporation, 100 Beaver St., Waltham, MA 02453-8443. Phone: (781) 398-2313. Fax: (781) 398-2476. E-mail: jerry.vovis@genomecorp.com.

TABLE 1. Gene targets of widely used antibiotics

Target category and gene product	Antibiotic class
Protein synthesis	
30S ribosomal subunit	Aminoglycosides, tetracyclines
50S ribosomal subunit	Macrolides, chloramphenicol
$tRNA^{Ile}$ synthetase	Mupirocin
Elongation factor G	Fusidic acid
Nucleic acid synthesis	
DNA gyrase A subunit; topoisomerase IV	Quinolones
DNA gyrase B subunit	Novobiocin
RNA polymerase beta subunit	Rifampin
DNA	Metronidazole
Cell wall peptidoglycan synthesis	
Transpeptidases	Beta-lactams
D-Ala-D-Ala ligase substrate	Glycopeptides
Antimetabolites	
Dihydrofolate reductase	Trimethoprim
Dihydropteroate synthesis	Sulfonamides
Fatty acid synthesis	Isoniazid

sequence for a second strain of *Mycobacterium tuberculosis* will appear soon (Table 2).

COMPARATIVE GENOMICS TO ASSESS THE SPECTRUM AND SELECTIVITY OF A TARGET

One powerful use of genomic sequence information is to compare all of the identified genes in different bacterial pathogens to determine which genes are, or are not, shared by various species. Indeed, Tatusov et al. (50) have suggested that gene families conserved among bacteria but missing from eukaryotes comprise a pool of potential targets for broad-spectrum antibiotic development. An early step in this direction was taken by Mushegian and Koonin (36), who identified 256 genes shared by the two completely sequenced bacterial genomes at that time, those of *Haemophilus influenzae* and *Mycoplasma genitalium*. On the other hand, genes which are apparently unique to a species such as *H. pylori* might be ideal for targeting that species with a narrow-spectrum antibiotic. As the number of sequenced bacterial and fungal genomes grows, so does the ability to find genes common to most microbial pathogens or truly unique to a particular species. For example, Arigoni et al. (6) identified 26 genes in *E. coli*, most of which were conserved in the *B. subtilis*, *M. genitalium*, *H. influenzae*, *H. pylori*, *Streptococcus pneumoniae*, and *Borrelia burgdorferi* genomes. They reasoned that this list of genes, which had no predictable function, contained novel targets for broad-spectrum antibiotic development. These analyses can be extended by including sequence comparisons to eukaryotic genomes as a means to examine potential selectivity of a target (50). For example, Arigoni et al. (6) reported that 15 of 26 proteins broadly conserved across bacterial species also exhibited significant sequence similarity to proteins in *S. cerevisiae* and, therefore, represented targets which, in an assay, might identify compounds that also have human toxicity. While these targets could simply be avoided, it should be noted that the targets of the majority of marketed antimicrobial agents show some conservation with mammalian proteins.

As in all sequence comparisons, the search parameters and the quality of the input data, e.g., partial human or mammalian sequence information, are critical. Relevant issues which must be addressed include questions such as the following. What degree of sequence similarity to another bacterial genome indicates a shared gene? What degree of sequence similarity to a mammalian gene warns of a possible toxicity problem? Since sequence similarity-searching algorithms allow nearly complete flexibility in the choice of these parameters, some known examples are necessary to calibrate the method. Mushegian and Koonin (36) used a BLASTP score of 90 as the cutoff for defining a biologically relevant relationship between two protein sequences. The appropriate cutoff score for exclusion of genes with apparent mammalian homologs may be more gene specific. Some examples reveal a general trend. Trimethoprim is a highly selective inhibitor of bacterial dihydrofolate reductases (DHFR) despite the fact that the human and *E. coli* DHFR gene products share 28% amino acid identity over the length of the two proteins (40). Similarly, the quinolones are highly selective against bacterial gyrases despite the fact that the C-terminal domain of human topoisomerase II shares 20% amino acid identity with *E. coli* gyrase A (25). Fluconazoles are highly selective for fungal lanosterol 14-α demethylases, even though the human and yeast gene products share 37% amino acid identity over their full length (5). These sequence identity percentages translate into BLASTP scores of 132, 125, and 301, respectively, in a search of a large nonredundant protein database comprised of sequences from GenBank, SwissProt, and PIR. Therefore, exclusion of genes having apparent mammalian homologs with scores >150 would likely be suitable for a search of bacterial targets, but the score cutoff would have to be raised to allow identification of the broadest set of antifungal target genes.

IDENTIFICATION OF ESSENTIAL TARGETS EXPERIMENTALLY

Genomic sequence information is not required for discovering essential genes, but such information does facilitate the process. Genes which are essential to pathogenesis and prevent

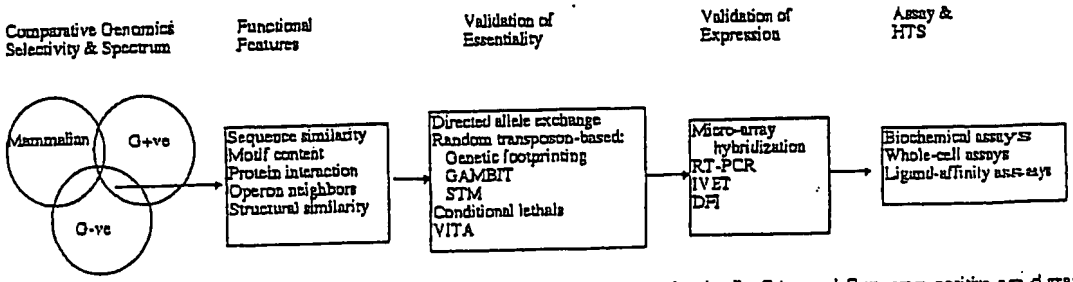


FIG. 1. Schematic view of genomic tools applied to antimicrobial-drug discovery. See the text for details. G+ve and G-ve, gram positive and gram negative, respectively.

TABLE 2. Sequenced microbial genomes

Internet resource	Genome	Strain(s)	Size (Mb)	Institution(s)	Reference
www.tigr.org/tdb/mdb/hidb/hidb.html	*Haemophilus influenzae* RD	KW20	1.83	TIGR	13
www.tigr.org/tdb/mdb/mgdb/mgdb.html	*Mycoplasma genitalium*	G-37	0.58	TIGR	15
www.tigr.org/tdb/mdb/mjdb/mjdb.html	*Methanococcus jannaschii*	DSM 2661	1.66	TIGR	8
www.kazusa.or.jp/cyano/cyano.html	*Synechocystis* sp.	PCC 6803	3.57	Kazusa DNA Research Institute	27
www.zmbh.uni-heidelberg.de/M_pneumoniae/MP_Home.html	*Mycoplasma pneumoniae*	M129	0.81	University of Heidelberg	23
speedy.mips.biochem.mpg.de/mips/yeast/yeast_genome.html or genome-www.stanford.edu/Saccharomyces	*Saccharomyces cerevisiae*	S288C	13	European and North American Consortium	17
www.tigr.org/tdb/mdb/hpdb/hpdb.html	*Helicobacter pylori*	26695	1.66	TIGR	51
www.genetics.wisc.edu/	*Escherichia coli*	K-12	4.6	University of Wisconsin	7
www.genomecorp.com/gene/sequences/methanobacter/abstract.html	*Methanobacterium thermoautotrophicum*	delta H	1.75	Genome Therapeutics and Ohio State University	43
www.pasteur.fr/Bio/SubtiList.html	*Bacillus subtilis*	168	4.2	International Consortium	31
www.tigr.org/tdb/mdb/afdb/afdb.html	*Archaeoglobus fulgidus*	VC-16, DSM4304	2.18	TIGR	29
www.tigr.org/tdb/mdb/bbdb/bbdb.html	*Borrelia burgdorferi*	B31	1.44	TIGR	14
www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=Genome&gi=133	*Aquifex aeolicus*	VF5	1.55	Diversa	10
www.bio.nite.go.jp/ot3db_index.html	*Pyrococcus horikoshii*	OT3	1.80	National Institute of Technology and Evaluation	28
www.sanger.ac.uk/Projects/M_tuberculosis/	*Mycobacterium tuberculosis*	H37Rv	4.40	Sanger Centre	9
www.tigr.org/tdb/mdb/tpdb/tpdb.html	*Treponema pallidum*	Nichols	1.14	TIGR and University of Texas	16
chlamydia-www.berkeley.edu:4231/	*Chlamydia trachomatis*	Serovar D (D/UW3/Cx)	1.05	University of California at Berkeley and Stanford University	46
evolution.bmc.uu.se/~siv/genomics/Rickettsia.html	*Rickettsia prowazekii*	Madrid E	1.11	University of Uppsala	4
www.genomecorp.com/hpylori or www.astro-boston.com/hpylori	*Helicobacter pylori*	J99	1.64	Genome Therapeutics and Astra AB	3
www.tigr.org/cig-bin/BlastSearch/blast.cgi?organism=m_tuberculosis	*Mycobacterium tuberculosis*	CSU#93	4.40	TIGR	Unpublished

colony formation in a conditional-lethal manner are potential targets for new antimicrobials. This assumes that a small organic molecule which inhibits the activity of an essential gene product would either kill or inhibit the growth of the bacterium which requires that functional protein. Such conditional lethal genes can be discovered through classical mutagenesis techniques. Availability of the sequence of the genome means that the full sequence of each mutated gene, and frequently its cellular role as well, can be gleaned from a short sequence read on a complementing plasmid insert. This additional information accelerates the processing of a mutational study enormously. Depending on the availability of genetic tools for the microbial species in question, a variety of molecular genetic methods can be used to discover essential genes. For example, in *E. coli*, genes can be placed under control of a regulated promoter by use of an appropriately constructed transposon system (11), or genes can be mutated to a conditional-lethal form. In principle, such conditional mutants can be used in whole-cell screens under moderately suppressing conditions in which the cells may be hypersensitive to drug-like compounds which act against that gene product (see below).

It seems reasonable to assume that most genes which are essential to the cell for growth or viability on laboratory media will also be required for growth or viability in an infected host. Experimentally, media can be varied in order to identify genes which are essential under the widest range of growth conditions and particularly in rich media which may simulate conditions in necrotic tissue of an animal host. Cells carrying auxotrophic mutations may find sufficient nutritional supplement in the host tissues to permit growth or at least survival. Such genes might be poor targets for new antimicrobials unless experiments establish that the particular nutrient is in short supply in the host or that cells are incapable of transporting the nutrient efficiently. In order to establish that a gene target is essential in an infection, a transposon-based gene tagging method called "signature-tagged mutagenesis" (STM) has been used to identify genes which are essential in an animal model (22, 35). However, since cells carrying the disrupted tagged genes must be grown in the laboratory prior to introduction into the animal, the method may be biased against genes which are essential for growth both on laboratory media and in an animal model. Indeed, many of the genes identified by STM appear to encode virulence factors which affect the ability of the pathogen to colonize or damage host tissue rather than the viability of the pathogen. New drugs which intervene in these processes could prove highly selective, and resistance to such drugs might be rare since loss or mutation of the virulence factor would also likely reduce virulence. However, other resistance mechanisms, such as drug modification and efflux pumps, could be problematic. In addition, the absence of a convenient in vitro assay for such drugs would hamper the development, testing, and approval processes. It remains unclear how many important antimicrobial targets would be missed by using as targets for drug discovery only those genes which are essential for growth or viability on laboratory culture media.

A related, important feature of a suitable antimicrobial gene target is its expression pattern in the infection. The absolute level of expression may be less important than information about whether it is expressed at all. A highly expressed, abundant gene product should be no more difficult to inhibit than a low-abundance gene product since an inhibitor with suitably high affinity will be effective in either case unless it is poorly taken up by pathogens. However, if a gene is not expressed at all in an established infection of an animal host, then it will be of no interest as a potential target. A gene already established as being essential for growth or viability in the laboratory by genetic methods obviously must be expressed under these conditions because its failure to be expressed as an active product causes the pathogen to die. Knowledge that such an essential

gene is also expressed in an animal model would suggest that it is essential in an infection as well. Two types of methods offer information about gene expression. First, for genes whose sequence is known, reverse transcriptase PCR (RT-PCR) may be used to detect transcripts in cells grown on agar media or in animal infection models (47). Alternatively, for organisms which have been sequenced in their entirety, a whole-genome view of gene expression may be obtained by gridding clones, PCR products, or synthetic oligonucleotides representing every gene onto a solid support. Total RNA may be isolated from cells grown under conditions of interest, labeled, and hybridized to the array (12). While thorough, this type of method suffers from some problems: (i) appropriate controls must be run to eliminate the possibility of bacterial DNA contamination in the RNA preparation, (ii) probes are difficult to prepare because bacterial mRNA is notoriously unstable, and (iii) the whole-genomic scale of the experiments makes the arrayed membranes difficult and expensive to prepare and read. A genetic promoter trap method termed "in vivo expression technology" or IVET may be more feasible for most laboratories (21, 33). In this approach, which has been developed for use in *Salmonella typhimurium* grown intraperitoneally in BALB/c mice or in cultured macrophages, random DNA fragments are cloned upstream from a gene whose expression is required for growth in an animal host. Cells, which multiply in vivo, are recovered and cloned. The sequences of fragments serving as functional promoters in vivo are then determined. A second, related promoter trap method termed "differential fluorescence induction" (DFI) has been described recently (53). The distinguishing features of this approach are that (i) the gene used for selection encodes a modified green fluorescent protein and (ii) the selection is accomplished with a fluorescence-activated cell sorter. If such methods can be extended to other bacterial species and animal hosts, they will be extremely useful for assessing random genomic fragments or specific genes of interest for expression in vivo.

IDENTIFICATION OF ESSENTIAL TARGETS USING DATABASES

Potential gene targets selected from databases can be validated by examining the effect of a gene knockout on cell growth or viability. Recombination is almost exclusively between homologous regions in bacterial genomes, and many common pathogens as well as model bacteria are transformable. Exchange between the chromosomal wild-type allele and a version engineered to carry a deletion and/or an insertion of a drug resistance cassette is generally efficient enough to be practical in the laboratory. Interpreting the results of such an experiment, however, may be difficult for two reasons. First, the frequent occurrence of polycistronic messages in bacteria means that disruption of a gene may have a deleterious effect on expression of a distal neighboring gene, a so-called "polar" effect. In that case, the inviability caused by a gene knockout could be due to loss of expression of a gene other than the one disrupted. Precautions can be taken to reduce these effects by, for example, including a moderate-strength outward reading promoter in the disrupted version of the allele so as to permit expression of the downstream gene(s). Second, the method works better as an exclusionary tool than as an inclusionary one. While success in generating a cell carrying a disrupted allele indicates that the gene is not essential for growth or viability of the cell, failure to generate such an altered cell could be due to any one of multiple causes including polar effects or inefficient recombination in a particular genetic interval.

TABLE 3. Additional Internet resources

Database or organization	Internet address
Sequence databases	
NCBI	http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
DDBJ	http://www.ddbj.nig.ac.jp/htmls_test/Welcome-e.html
EBI/EMBL	http://www.ebi.ac.uk/ebi_home.html
GSDB	http://www.ncgr.org/gsdb/Index_gsdb.html
SwissProt (Geneva)	http://expasy.hcuge.ch/www/expasy-top.html
Candida	http://alces.med.umn.edu/Candida.html
MIPs	http://www.mips.biochem.mpg.de/
RDP	http://rdp.life.uiuc.edu/
SGD	http://genome-www.stanford.edu/
Metabolic databases	
KEGG	http://www.genome.ad.jp/kegg/
Ecocyc	http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html
WIT	http://www.cme.msu.edu/WIT/
Sequencing groups	
Berkeley	http://chlamydia-www.berkeley.edu:4231/
Genome Therapeutics	http://www.genomecorp.com/home.htl
Sanger	http://www.sanger.ac.uk/Projects/
Stanford	http://sequence-www.stanford.edu/group/malaria/index.html
TIGR	http://www.tigr.org/tdb/mdb/mdb.html
University of Oklahoma	http://dna1.chem.uoknor.edu/index.html
University of Queensland	http://www.cmcb.uq.edu.au/aeruginosa/
University of Washington	http://chimera.biotech.washington.edu/uwgc/
Washington University	http://genome.wustl.edu/gsc/bacterial/salmonella.html
Tools and resources	
Biomolecular Research Tools	http://www.public.iastate.edu/~pedro/rt_1.html
COGs	http://www.ncbi.nlm.nih.gov/COG/
NCGR	http://www.ncgr.org/microbe/Index_bns.html
MAGPIE	http://www.mcs.anl.gov/home/gunsteri/genomes.html
Genobase	http://specter.dcrt.nih.gov:8004/
Micro Underground	http://www.lsume.edu/campus/micr/mirror/public_html/index.html
ANMR	http://www.wdcm.riken.go.jp/
WHO	http://www.who.ch/Welcome.html
Pallen	http://www.qmw.ac.uk/~rh bm001/ketbook/chapter.html
CDC	http://www.cdc.gov/
University of Kansas	http://www.kumc.edu/research/fgsc/main.html
University of Georgia	http://fungus.genetics.uga.edu:5080/
Tripos	http://www.tripos.com/sites.html
Motif	http://dna.Stanford.EDU/identify/
Pedant	http://pedant.mips.biochem.mpg.de/
GenTHREADER	http://globin.bio.warwick.ac.uk/genome/genomic.html

One solution to this problem is to carry out allele exchange as a two-step process (20, 32). In *E. coli*, for example, the disrupted allele together with the vector carrying it can be integrated into the genome by means of a single crossover, a so-called "Campbell insertion." Recombination between homologous regions on the two copies of the allele now on the chromosome will eliminate the vector sequences and either copy of the allele. Which copy is eliminated depends upon which regions of homology were involved in the recombination. Failure to find cells retaining only the disrupted allele strongly suggests that such progeny are inviable. Success in finding cells retaining only the wild-type allele confirms that

recombination is efficient in this genetic interval. However, in many naturally competent bacterial species, such as *H. influenzae* and *S. pneumoniae*, double-crossover events are extremely efficient, and allele replacement occurs with little or no opportunity to isolate a single crossover intermediate (1). While this complicates evaluation of essential genes in these organisms, it provides a convenient method for disrupting genes under conditions in which they are not essential so that the resulting strains may be examined under a variety of other conditions (e.g., see below).

A new approach promises to accelerate the process of evaluating the essentiality of genes. Smith et al. (44, 45) have described a method for the yeast *S. cerevisiae* called "genetic footprinting" which makes use of a quasi-random transposable Ty element to generate a rich array of gene knockouts in a population of cells. Further transposition is shut off, and the population is then grown under a variety of conditions. DNA is prepared from cells in the various growth populations, and the DNA is queried by PCR amplification to determine if it will yield PCR products between a gene-specific primer and a transposon-specific primer. Failure to find such PCR products suggests that cells carrying transposons in that gene were inviable under the growth conditions employed. Fluorescent PCR products are viewed on standard sequencing gels by using automated fluorescence sequencing machines and a commercially available software package. An important control in this method is the existence of a gene-to-transposon PCR product in the so-called t_0 cell population prior to the shutdown of transposition. This assures the experimenter that this region is not simply a "cold" spot for transposition. The efficiency of this method derives from the use of random transposons to build all necessary gene knockouts rapidly, followed by automated PCR and analysis methods to interpret the results for any given gene of interest.

Recently, a modified version of this method, called "genomic analysis and mapping by in vitro transposition" (GAMBIT), has been applied successfully to two bacterial species (1). In this variation of genetic footprinting, the transposition mutagenesis was done on PCR-amplified genomic segments from *H. influenzae* or *S. pneumoniae* in vitro, and the mutations were introduced into these naturally competent host bacteria by transformation. While the method suffers from the absence of a true t_0, the focus on 10-kb DNA segments permits near-saturation mutagenesis with the *mariner* family transposon *Himar1*, which shows little or no insertion site specificity. These authors identified four essential conserved genes of unknown function from a total of 13 analyzed.

Currently, the main limitation to this method is a requirement for an efficiently transformable host bacterium so that mutations generated in vitro can be evaluated readily in vivo. Other limitations which apply to all genetic footprinting methods include the following: (i) essentiality of the function of a gene that is duplicated or has a functional paralog cannot be analyzed, since footprinting assesses the fitness of a single mutagenized gene; (ii) polarity effects, although not a problem for *S. cerevisiae*, may lead to misinterpretation of data obtained from bacteria; (iii) the correlation of footprinting data with gene knockout data has not been confirmed in any organism; and (iv) footprinting data are technically difficult to interpret for a variety of reasons, including the facts that some essential genes will tolerate insertions in the C-terminal coding region (e.g., *secA* [1]) and cells carrying insertions in some genes display an intermediate slow-growth phenotype (e.g., *ade2* [44]).

TOOLS FOR PREDICTING THE FUNCTION OF GENE PRODUCTS

Clearly, not all of the predicted functional assignments based on sequence similarities are reliable. In some cases, for example, the function of the closest-related protein has itself been predicted based on its sequence similarity to a gene product of known function. In other cases, the chain of relatedness to a protein of confirmed function may be even longer. About half of the genes in bacterial genomes either lack significant enough sequence similarity to permit functional assignment or have likely homologs whose function is unknown. In neither of these cases can a function be predicted for the gene product. Nevertheless, the results of sequence similarity searches are a useful starting point for further investigation. More sensitive sequence comparison searches may provide a putative function or functional feature such as the presence of a short protein sequence motif. For example, a search against a database of clusters of orthologous groups of genes (COGs [Table 3]) yielded over 100 additional functional predictions for genes in the *H. pylori* genome (50).

Tools other than sequence similarity have also been useful in a few cases for predicting function of a gene product. For example, a gene product, with no significant sequence relationship to a protein of known function but which is likely to be cotranscribed as part of a polycistronic message with other genes of known function, may play a role in the same pathway with the known gene products. In the *E. coli* genome, the hypothetical gene *yjaF* appears to be cotranscribed with the porphyrin biosynthetic gene *hemE*, and the hypothetical gene *yadM* appears to be in an operon with the outer-membrane usher protein HtrE, which is involved in transport and binding. It is reasonable to speculate that these genes of unknown function play roles in the same biochemical pathways as their neighboring "known" genes. Of course, experimental evidence would be required to confirm these hypotheses. Methods also exist for identifying likely structural similarity even in the absence of strong primary sequence similarity. As the databases of known structures grow, this will become a powerful approach for assigning likely functions to gene products. For example, the "GenTHREADER" web site (Table 3) presents analysis results from a fast fold recognition program on the predicted open reading frames from three bacterial genomes.

Laboratory methods can also be invoked to solve questions of unknown gene identities. An unknown gene may be used as the bait in a yeast two-hybrid interaction trap to identify genes whose protein products interact with the unknown protein. The identity of an interacting partner will frequently implicate the unknown in a particular cellular pathway (19). Finally, an unknown gene may be expressed as a tagged fusion, the protein purified by affinity column, and the product tested for categories of activities such as proteolysis, DNA cleavage or binding, ATP or GTP hydrolysis, and binding, to name a few. The probability of successfully identifying an activity of an unknown by the latter method is low, but this method may be warranted if sequence comparisons suggest the presence of a motif associated with an assayable function. An attractive alternative is to focus on assays which do not require knowledge of the cellular function of a gene product (see below).

THE FUTURE: DEALING WITH GENE TARGETS HAVING NO PREDICTABLE FUNCTIONAL FEATURES

The array of tools described so far, including comparative genomic methods for identifying potentially useful gene targets and allele exchange methods for validating the essentiality of

those genes, provides both gene targets whose cellular function can be predicted and gene targets for which little or no functional information is available. Targets in the first class may be used immediately to build biochemical assays and high-throughput screens to detect small organic molecules which inhibit the biochemical activity. Typically, the gene sequence is amplified by PCR from genomic DNA of a given bacterium, inserted into an expression vector, and expressed in *E. coli* sometimes with affinity tags to facilitate purification of the resulting protein product.

It is far less obvious how to proceed with gene targets lacking any functional information. This problem has attracted considerable attention in recent years because of the growing number of such targets known to be shared across many bacterial species (24), some of which are known to be essential in at least one species. As a general guide, about 40% of bacterial genes cannot be assigned a putative function at this time. If 10 to 15% of these genes are essential, then 4 to 6% of the genes in a typical bacterial genome (about 100 genes) represent potential antimicrobial targets which have never been used in screens. Three basic types of approaches seem feasible and have shared some initial success. First, cells expressing higher- or lower-than-normal levels of particular genes have in some cases been shown to be more resistant or more sensitive, respectively, than their wild-type parents to chemical compounds known to inhibit those gene products. For example, overexpression of the yeast *ALG7* gene results in cells more resistant than wild-type cells to tunicamycin (38), while reduced activity of the same gene product results in cells more sensitive to the drug (30). Similarly, increased expression of the *ERG11* gene in *Candida glabrata* results in higher levels of resistance to the azole family of drugs which target that enzyme (54). A gene of unknown function could be overexpressed in a host strain, and the resulting assay strain could be tested for increased resistance to a library of compounds. It is clear, however, that many gene targets when overexpressed do not lead to resistance to chemical compounds that are known to bind to the protein product (e.g., *gyrA* [52]). Furthermore, overexpression of proteins often leads to lethality or growth defects (e.g., *kasA* [34]). Alternatively, a gene could be underexpressed or crippled by a mutation so that cells might show increased sensitivity to a compound which inhibits the protein product. Scientists at Microcide Pharmaceuticals, Inc., have applied this approach on a large scale using temperature-sensitive mutants grown at intermediate temperatures in order to reduce the level of activity of the target gene product (39a). Of course, it is not clear what fraction of unknown gene products would provide the cell with increased drug resistance or sensitivity when over- or underexpressed in these ways.

The second approach to this problem of assaying gene products of unknown function is probably more generally applicable. Libraries of small molecules are screened for strong binding affinity to proteins of unknown function. This has been achieved with peptides in phage display libraries because binding can be readily detected by elution of bound phage from the protein tethered on a solid support. Proteins of unknown function can be produced easily as affinity fusion products for attachment to solid supports, and a variety of peptide phage display libraries are commercially available. Conformationally constrained disulfide-bonded peptides with affinities in the 100 μM to 100 nM range can be obtained by this approach (55). Of course, not all peptides detected by this approach will bind to sites which inhibit activity, but an elegant new method, called "validation in vivo of targets for anti-infectives" (VITA), has been devised to identify those peptides which inhibit essential cellular functions (49). Potential inhibitory peptides were expressed in a regulated manner within bacterial host cells which were grown either on agar medium or in an animal model of infection. Inhibition of cell growth or viability upon induction of peptide expression validated the peptide-protein interaction as useful for further drug development. While peptides are not ideal drug candidates, a wider array of techniques are applicable after a moderate binder has been obtained. The peptide may be used as a surrogate ligand in a competition assay to identify a small organic compound with higher affinity. Scintillation proximity assays (26) or fluorescence polarization assays (41) may be used in a high-throughput mode to identify compounds in chemical libraries which compete for binding with a labeled peptide. Alternatively, ligand binding assays may be configured to work directly on libraries of unlabeled chemical compounds. Shuker et al. (42) have described a nuclear magnetic resonance-based method capable of a throughput of 1,000 compounds per day. Mass spectrometric methods are also of interest as potentially rapid ways to detect bound ligands from chemical libraries. One concern about these approaches is that proteins may have multiple accessible binding sites, many of which have nothing to do with catalytic activity. It is not clear at this early stage how significant an issue multiple binding sites will be. However, it is worth noting that Shuker et al. (42) took advantage of a second binding site to increase the affinity of an inhibitor for the protein. Ultimately, of course, affinity ligands must be shown to inhibit cell growth, that is, to have antimicrobial activity. Some chemical engineering of the compound may be required to increase microbial uptake.

A third approach for assaying gene products of unknown function relies on the complex gene expression regulatory network found in many bacteria. Expression levels of genes in metabolic pathways are often regulated in response to the amounts of intermediates in the cell. For example, disruption of the general secretory pathway in *E. coli* by mutation results in dramatic up-regulation of *secA* gene expression (37). Alksne et al. (2) took advantage of this fact to build a strain of *E. coli* carrying a *secA-lacZ* fusion as a detectable reporter. Several synthetic compounds and natural products were identified by their ability to induce expression of the reporter. Many of these exhibited antimicrobial activity and reduced the secretion of *Staphylococcus aureus* toxin 1. Similarly, Mdluli et al. (34) have reported that sublethal concentrations of isoniazid lead to up-regulation of the *kasA* and *acpM* genes. This group has initiated a whole-cell, high-throughput screen of chemical compounds which induce expression of a luciferase reporter fused to a gene in this regulated pathway. Screens of this type, which take advantage of the bacterial gene regulatory network, are inherently less specific than the two other types described here. In addition, they suffer from the basic limitation of all whole-cell screens: compounds must be capable of entering the cell in order to be detected. However, these types of screens offer the potential advantage of identifying compounds which act at any of several points in a pathway.

CONCLUSIONS

The availability of genomic sequence information for all or nearly all of several different bacterial species provides important new advantages for target discovery. First, it permits use of a comparative genomic analysis to identify potential new targets shared across several bacterial species or particular to a single species. In this manner, it is possible to generate lists of genes which represent potential targets for broad-spectrum or highly focused narrow-spectrum antibiotics. Sequence comparisons can also provide some assurance against mammalian

toxicity if proteins of similar sequence do not exist in mammalian sequence databases. Second, sequence similarity provides some insights into putative functions for most gene products. Finally, availability of the entire sequence of the gene target of interest permits rapid construction of gene knockouts to validate the utility of the target and facile construction of expression plasmids for production of protein and development of assays. The fact that bacterial and fungal genes can be assessed rapidly for their relevance as potential antibiotic targets by determining the effect of knocking out the gene and the fact that their genomes are small enough to be sequenced in their entirety are compelling reasons that the field of genomics will likely find its first real utility in the development of new antimicrobials.

ACKNOWLEDGMENTS

We thank our colleagues at Genome Therapeutics Corporation and the Schering-Plough Research Institute for helpful discussions about genomic approaches to drug discovery. In particular, Skip Shimer, Brad Guild, and Lucy Ling were instrumental in the analysis of the approaches summarized here. We thank Douglas Smith of Genome Therapeutics Corporation for the compilation of Internet resources presented in Table 3.

REFERENCES

1. Akerley, B. J., E. J. Rubin, A. Camilli, D. J. Lampe, H. M. Robertson, and J. J. Mekalanos. 1998. Systematic identification of essential genes by *in vitro mariner* mutagenesis. Proc. Natl. Acad. Sci. USA 95:8927–8932.
2. Alksne, L. E., P. Burgio, P. Bradford, B. Feld, W. Hu, P. Labthavikul, M. McGlynn, P. J. Petersen, M. Tuckman, and S. Projan. 1998. Identification of inhibitors of bacterial secretion by using a SecA reporter system, p. 272. *In* Abstracts of the 38th Interscience Conference on Antimicrobial Agents and Chemotherapy. American Society for Microbiology, Washington, D.C.
3. Alm, R. A., L. L. Ling, D. T. Moir, B. L. King, E. D. Brown, P. C. Doig, D. R. Smith, B. Noonan, B. C. Guild, B. L. deJonge, G. Carmel, P. J. Tummino, A. Caruso, M. Uria-Nickelsen, D. M. Mills, C. Ives, R. Gibson, D. Merberg, S. D. Mills, Q. Jiang, D. E. Taylor, G. F. Vovis, and T. J. Trust. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397:176–180.
4. Andersson, S. G. E., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, U. C. M. Alsmark, R. M. Podowski, A. K. Naslund, A.-S. Eriksson, H. H. Winkler, and C. G. Kurland. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140.
5. Aoki, Y., F. Yoshihara, M. Kondoh, Y. Nakamura, N. Nakayama, and M. Arisawa. 1993. Ro 09-1470 is a selective inhibitor of P-450 lanosterol C-14 demethylase of fungi. Antimicrob. Agents Chemother. 37:2662–2667.
6. Arigoni, F., F. Talabot, M. Peitsch, M. D. Edgerton, E. Meldrum, E. Allet, R. Fish, T. Jamotte, M.-L. Curchod, and H. Loferer. 1998. A genome-based approach for the identification of essential bacterial genes. Nat. Biotechnol. 16:851–856.
7. Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1462.
8. Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, and J. C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073.
9. Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–544.
10. Deckert, G., P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox, D. E. Graham, R. Overbeek, M. A. Snead, M. Keller, M. Aujay, R. Huberk, R. A. Feldman, J. M. Short, G. J. Olsen, and R. V. Swanson. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 393: 353–358.
11. de Lorenzo, V., L. Eltis, B. Kessler, and K. N. Timmis. 1993. Analysis of *Pseudomonas* gene products using *lacI*^{q}/*Ptrp-lac* plasmids and transposons that confer conditional phenotypes. Gene 123:17–24.
12. DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686.
13. Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512.
14. Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, M. Gwinn, B. Dougherty, J.-F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, A. R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M. D. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fujii, M. D. Cotton, K. Horst, K. Roberts, B. Hatch, H. O. Smith, and J. C. Venter. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390: 580–586.
15. Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison III, and J. C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403.
16. Fraser, C. M., S. J. Norris, G. M. Weinstock, O. White, G. G. Sutton, R. Dodson, M. Gwinn, E. K. Hickey, R. Clayton, K. A. Ketchum, E. Sodergren, J. M. Hardham, M. P. McLeod, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, J. K. Howell, M. Chidambaram, T. Utterback, L. McDonald, P. Artiach, C. Bowman, M. D. Cotton, J. C. Venter, et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388.
17. Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. 1996. Life with 6000 genes. Science 274:546–567.
18. Gold, H. S., and R. C. Moellering. 1996. Antimicrobial-drug resistance. N. Engl. J. Med. 335:1445–1453.
19. Gyuris, J., E. Golemis, H. Chertkov, and R. Brent. 1993. Cdi1, a human G1 and S phase protein phosphatase that associates with Cdk2. Cell 75:791–803.
20. Hamilton, C. M., M. Aldea, B. K. Washburn, P. Babitzke, and S. R. Kushner. 1989. New method for generating deletions and gene replacements in *Escherichia coli*. J. Bacteriol. 171:4617–4622.
21. Helfhoff, D. M., C. P. Conner, P. C. Hanna, S. M. Julio, U. Hentschel, and M. J. Mahan. 1997. Bacterial infection as assessed by *in vivo* gene expression. Proc. Natl. Acad. Sci. USA 94:934–939.
22. Hensel, M., J. E. Shea, C. Gleeson, M. D. Jones, E. Dalton, and D. W. Holden. 1995. Simultaneous identification of bacterial virulence genes by negative selection. Science 269:400–403.
23. Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 24:4420–4449.
24. Hinton, J. C. D. 1997. The *Escherichia coli* genome sequence: the end of an era or the start of the FUN? Mol. Microbiol. 26:417–422.
25. Hoshino, K., K. Sato, T. Une, and Y. Osada. 1989. Inhibitory effects of quinolones on DNA gyrase of *Escherichia coli* and topoisomerase II of fetal calf thymus. Antimicrob. Agents Chemother. 33:1816–1818.
26. Jenh, C. H., M. Zhang, M. Wiskowski, J. C. Tan, X. D. Fan, V. Hegde, M. Patel, R. Bryant, S. K. Narula, P. J. Zavodny, and C. C. Chou. 1998. Development of a CD28 receptor binding-based screen and identification of a biologically active inhibitor. Anal. Biochem. 256:47–55.
27. Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, and S. Tabata. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3: 109–136.
28. Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, and H. Kikuchi. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res. 5(Suppl.):147–155.
29. Klenk, H.-P., R. A. Clayton, J.-F. Tomb, O. White, K. E. Nelson, K. A.

Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, D. L. Richardson, A. R. Kerlavage, D. E. Graham, N. C. Kyrpides, R. D. Fleischmann, J. Quackenbush, N. H. Lee, G. G. Sutton, S. Gill, E. F. Kirkness, B. A. Dougherty, K. McKenney, M. D. Adams, B. Loftus, S. Peterson, C. I. Reich, L. K. McNeil, J. H. Badger, A. Glodek, L. Zhou, R. Overbeek, J. D. Gocayne, J. F. Weidman, L. McDonald, T. Utterback, M. D. Cotton, T. Spriggs, P. Artiach, B. P. Kaine, S. M. Sykes, P. W. Sadow, K. P. D'Andrea, C. Bowman, C. Fujii, S. A. Garland, T. M. Mason, G. J. Olsen, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390:364–370.

30. Kukuruzinska, M. A., and K. Lennon. 1995. Diminished activity of the first N-glycosylation enzyme, dolichol-P-dependent N-acetylglucosamine-1-P transferase (GPT), gives rise to mutant phenotypes in yeast. Biochim. Biophys. Acta 1247:51–59.
31. Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessières, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S.-K. Choi, J.-J. Codani, I. F. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, K. M. Devine, A. Düsterhöft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S.-Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, et al. 1997. The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. Nature 390:249–256.
32. Link, A. J., D. Phillips, and G. M. Church. 1997. Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization. J. Bacteriol. 179:6228–6237.
33. Mahan, M. J., J. W. Tobias, J. M. Slauch, P. C. Hanna, R. J. Collier, and J. J. Mekalanos. 1995. Antibiotic-based selection for bacterial genes that are specifically induced during infection of a host. Proc. Natl. Acad. Sci. USA 92:669–673.
34. McNeil, K., R. A. Slayden, Y.-Q. Zhu, S. Ramaswamy, X. Pan, D. Mead, D. D. Crane, J. M. Musser, and C. E. Barry. 1998. Inhibition of a *Mycobacterium tuberculosis* β-ketoacyl ACP synthase by isoniazid. Science 280:1607–1610.
35. Mei, J. M., F. Nourbakhsh, C. W. Ford, and D. W. Holden. 1997. Identification of Staphylococcus aureus virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis. Mol. Microbiol. 26:399–407.
36. Mushegian, A. R., and E. V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. USA 93:10268–10273.
37. Riggs, P. D., A. I. Derman, and J. Beckwith. 1988. A mutation affecting the regulation of a secA-lacZ fusion defines a new sec gene. Genetics 118:571–579.
38. Rine, J. 1991. Gene overexpression in studies of *Saccharomyces cerevisiae*. Methods Enzymol. 194:239–251.
39. Salyers, A. A., and C. F. Amabile-Cuevas. 1997. Why are antibiotic resistance genes so resistant to elimination? Antimicrob. Agents Chemother. 41:2321–2325.
39a. Schmid, M. Personal communication.
40. Schweitzer, B. I., A. P. Dicker, and J. R. Bertino. 1990. Dihydrofolate reductase as a therapeutic target. FASEB J. 4:2441–2452.
41. Seethala, R., and R. Menzel. 1997. A homogeneous, fluorescence polarization assay for src-family tyrosine kinases. Anal. Biochem. 253:210–218.
42. Shuker, S. B., P. J. Hajduk, R. P. Meadows, and S. W. Fesik. 1996. Discovering high-affinity ligands for proteins: SAR by NMR. Science 274:1531–1534.
43. Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush, H. Safer, D. Patwell, S. Prabhakar, S. McDougall, G. Shimer, A. Goyal, S. Pietrokovski, G. M. Church, C. J. Daniels, J.-I. Mao, P. Rice, J. Noelling, and J. N. Reeve. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. J. Bacteriol. 179:7135–7155.
44. Smith, V., D. Botstein, and P. O. Brown. 1995. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. Proc. Natl. Acad. Sci. USA 92:6479–6483.
45. Smith, V., K. N. Chou, D. Lashkari, D. Botstein, and P. O. Brown. 1996. Functional analysis of the genes of yeast chromosome V by genetic footprinting. Science 274:2069–2074.
46. Stephens, R. S., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R. L. Tatusov, Q. Zhao, E. V. Koonin, and R. W. Davis. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754–759.
47. Swartley, J. S., L.-J. Liu, Y. K. Miller, L. E. Martin, S. Edupuganti, and D. S. Stephens. 1998. Characterization of the gene cassette required for biosynthesis of the (α1→6)-linked *N*-acetyl-D-mannosamine-1-phosphate capsule of serogroup A *Neisseria meningitidis*. J. Bacteriol. 180:1533–1539.
48. Swartz, M. N. 1994. Hospital-acquired infections: diseases with increasingly limited therapies. Proc. Natl. Acad. Sci. USA 91:2420–2427.
49. Tao, J., T. Li, G. Connelly, X. Shen, J. Silverman, F. Houman, P. Wendler, and F. P. Tally. 1998. VITA: validation in vivo of targets and assays for anti-infectives, p. 274. *In* Abstracts of the 38th Interscience Conference on Antimicrobial Agents and Chemotherapy. American Society for Microbiology, Washington, D.C.
50. Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. Science 278:631–637.
51. Tomb, J.-F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E. F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzegerald, N. Lee, M. D. Adams, E. K. Hickey, D. E. Berg, J. D. Gocayne, T. R. Utterback, J. D. Peterson, J. M. Kelley, M. D. Cotton, J. M. Weidman, C. Fujii, C. Bowman, L. Watthey, E. Wallin, W. S. Hayes, M. Borodovsky, P. D. Karp, H. O. Smith, C. M. Fraser, and J. C. Venter. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539–542.
52. Truong, Q. C., J. C. Nguyen Van, D. Shlaes, L. Gutmann, and N. J. Moreau. 1997. A novel, double mutation in DNA gyrase A of Escherichia coli conferring resistance to quinolone antibiotics. Antimicrob. Agents Chemother. 41:85–90.
53. Valdivia, R. H., and S. Falkow. 1997. Fluorescence-based isolation of bacterial genes expressed within host cells. Science 277:2007–2011.
54. Vandenbossche, H., P. Marichal, F. C. Odds, L. Lejeune, and M. C. Coene. 1992. Characterization of an azole-resistant *Candida glabrata* isolate. Antimicrob. Agents Chemother. 36:2602–2610.
55. Wrighton, N. C., F. X. Farrell, R. Chang, A. K. Kashyap, F. P. Barbone, L. S. Mulcahy, D. L. Johnson, R. W. Barrett, L. K. Jolliffe, and W. J. Dower. 1996. Small peptides as potent mimetics of the protein hormone erythropoietin. Science 273:458–463.

A Genomic Perspective on Protein Families

Roman L. Tatusov, Eugene V. Koonin,* David J. Lipman

In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Comparison of proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allowed the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This relation automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

The release in 1995 of the complete genome sequence of the bacterium *Haemophilus influenzae* (*1*), followed within the next 1.5 years by four more bacterial genomes (*2*), one archaeal genome (*3*), and one genome of a unicellular eukaryote (*4*), marked the advent of a new age in biology. The hallmark of this era is that comparisons between complete genomes are becoming an indispensable component of our understanding of a variety of biological phenomena. The number of sequenced genomes is expected to grow exponentially for at least the next few years, and conceivably, their impact on biology will further increase (*5*).

Knowing the inventory of conserved genes responsible for housekeeping functions and understanding the differences in the genetic basis of these functions in different phylogenetic lineages is central to understanding life itself, at least at the level of a single cell. Complete sequences are indispensable for achieving this goal because they hold the only type of information that can be used to delineate the complete network of relationships between genes from different genomes. Furthermore, only with complete genome sequences is it possible to ascertain that a particular protein implicated in an essential function is not encoded in a given genome. Accordingly, an alternative protein for the respective function should be sought among the functionally unassigned gene products (*6*). With multiple genome sequences, it is possible to delineate protein families that are highly conserved in one domain of life but are missing in the others. Such information may be critically important: For example, the families that are conserved among bacteria but are missing in eukaryotes comprise the pool of potential targets for broad-spectrum antibiotics.

The knowledge of all of the gene sequences from multiple complete genomes redefines the problem of gene classification. It becomes feasible to replace the more or less arbitrary clustering of genes by similarity with a complete, consistent system in which the groups are likely to have evolved from a single ancestral gene. Such a natural classification of genes will provide a framework for evolutionary studies and for rapid, largely automatic functional annotation of newly sequenced genomes. This framework will evolve and improve with increasing coverage of the diversity of life forms with complete genome sequences. It is critical to have this system in place while the number of completed genomes is still small and each family can be explored individually. Here we describe a prototype of a natural system of gene families from complete genomes.

Orthologs and Paralogs: Deriving Clusters of Orthologous Groups

The relationships between genes from different genomes are naturally represented as a system of homologous families that include both orthologs and paralogs. Orthologs are genes in different species that evolved from a common ancestral gene by speciation; by contrast, paralogs are genes related by duplication within a genome (*7*). Normally, orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if related to the original one. Thus, identification of orthologs is critical for reliable prediction of gene functions in newly sequenced genomes. It is equally important for phylogenetic analysis because interpretable phylogenetic trees generally can be constructed only within sets of orthologs (*8*). A complete list of orthologs also is a prerequisite for any meaningful comparison of genome organization (*9*).

A naïve operational definition would simply maintain that for a given gene from one genome, the gene from another genome with the highest sequence similarity is the ortholog. Given the complete genome sequences, this straightforward approach often gives credible results, especially when the compared species are not too distant phylogenetically (*9*). At larger phylogenetic distances, however, the situation becomes more complicated. If gene duplications occurred in each of the given two clades subsequent to their divergence, only a many-to-many relationship will adequately describe orthologs, and accordingly, detection of the highest similarity will not result in the identification of the complete set of orthologs. In addition, when the best hit is not highly significant statistically, which is common in the case of phylogenetically distant relationships (*10*), it simply may be spurious. On the other hand, attempts to apply a restrictive similarity cutoff are likely to result in a number of orthologs being missed.

Given the existence of one-to-many and many-to-many orthologous relationships, we redefined the task of identifying orthologs as the delineation of clusters of orthologous groups (COGs). Each COG consists of individual orthologous genes or orthologous groups of paralogs from three or more phylogenetic lineages. In other words, any two proteins from different lineages that belong to the same COG are orthologs. Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events.

In order to delineate the COGs, all pairwise sequence comparisons among the 17,967 proteins encoded in the seven complete genomes were performed (*11*), and for each protein, the best hit (BeT) in each of the other genomes was detected. The identification of COGs was based on consistent patterns in the graph of BeTs. The simplest and most important of such patterns is a triangle, which typically consists of orthologs (Fig. 1A). Indeed, if a gene from one of the compared genomes has BeTs in two other genomes, it is highly unlikely that the respective genes are also BeTs for one another unless they are bona fide orthologs (*12*). The consistency between BeTs resulting in triangles does not depend on the absolute level of similarity between the compared proteins and thus allows the detection of orthologs among both slowly and quickly evolving genes. This approach is most likely to be informative when the

The authors are with the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

*To whom requests for reprints should be addressed. E-mail: koonin@ncbi.nlm.nih.gov

Exhibit 2

BeTs forming a triangle come from widely different lineages. Accordingly, only five major, phylogenetically distant clades were used as independent contributors to COGs: Gram-negative bacteria (*Escherichia coli* and *H. influenzae*), Gram-positive bacteria (*Mycoplasma genitalium* and *M. pneumoniae*), Cyanobacteria (*Synechocystis* sp.), Archaea (Euryarchaeota) (*Methanococcus jannaschii*), and Eukarya (Fungi) (*Saccharomyces cerevisiae*) (*13*).

The procedure used to derive COGs included finding all triangles formed by BeTs between the five major clades and merging those triangles that had a common side until no new ones could be joined. A triangle is an elementary, minimal COG (Fig. 1A). The groups produced by merging adjacent triangles include orthologs from different lineages and, in many cases, paralogs from the same lineage (Fig. 1, B and C). Because of the existence of paralogs, the BeTs that form the triangles are not necessarily symmetrical: For example, in the COG shown in Fig. 1C, the same *M. genitalium* protein, MG249, is the BeT for four paralogous σ subunits of *E. coli* RNA polymerase, but only for one of them, RpoD, is the relationship symmetrical.

Most of the clusters derived by the above procedure meet the definition of a COG, that is, all of the proteins from the different lineages in the same cluster are likely to be orthologs. There are, however, several reasons why, in certain cases, COGs may be lumped together. Proteins may contain two or more distinct regions, each of which belongs to a different conserved family; usually such proteins are loosely referred to as multidomain (*14*). Each of the clusters was inspected for the presence of multidomain proteins, individual domains were isolated (*15*), and a second iteration of the sequence comparison was performed with the resulting database of domains. Some of the COGs may include proteins from different lineages that are paralogs rather than orthologs, primarily because of differential gene loss in the major phylogenetic lineages. When one gene in a pair of paralogs is lost in one lineage but not in the others, two COGs that should have been distinct may be artificially joined. Therefore, the level of sequence similarity between the members of each cluster was analyzed, and clusters that seemed to contain two or more COGs were split.

Phylogenetic and Functional Patterns in COGs

The described analysis resulted in 710 apparent COGs. This set appears to be essentially complete as far as orthologous relationships are concerned. Indeed, when the portion of the database of proteins from complete genomes not included in the COGs was clustered by sequence similarity (*16*), only 10 groups were identified, which, upon careful inspection of the alignments, were considered likely to constitute additional COGs missed originally. These groups were incorporated, producing the final collection of 720 COGs, including 6814 proteins and distinct domains of multidomain proteins (6646 distinct gene products, or 37% of the total number of genes in the seven complete genomes) (*17*).

Most of the COGs are relatively small groups of proteins. One-third of the COGs (240 COGs with 1406 proteins) contain one representative of each of the included species (no paralogs), and 192 more COGs include paralogs from only one species, most frequently yeast (87 COGs). The mean number of proteins per COG increases with increasing number of genes in a genome, from 1.2 for *M. genitalium* to 2.9 for yeast. A notable aspect of many COGs is the differential behavior of paralogs. It is typical that one of the paralogs, for example, in yeast, shows consistently higher similarity to the orthologs in all or most of the other species (Fig. 1, B and C). For numerous yeast paralogs, particularly components of the translation apparatus, the underlying cause is obvious: the gene whose product is most similar to the bacterial orthologs is of mitochondrial origin (Fig. 1B). A more common explanation for the asymmetry of the relationships in the COGs, however, is that the highly conserved paralog has retained the original function, whereas the functions of the less conserved paralogs have changed in the course of evolution. In the already considered example (Fig. 1C), the symmetrical component of the graph (solid lines) delineates the conserved function of the σ70 subunit of the RNA polymerase (*E. coli* RpoD), which is required for the transcription of the bulk of bacterial genes, whereas the asymmetrical BeTs (broken lines) are observed for σ subunits (*E. coli* RpoH, RpoS, and FliA) involved in the transcription of specialized gene subsets (*18*). This phenomenon appears to be widespread, as we found 549 proteins in 302

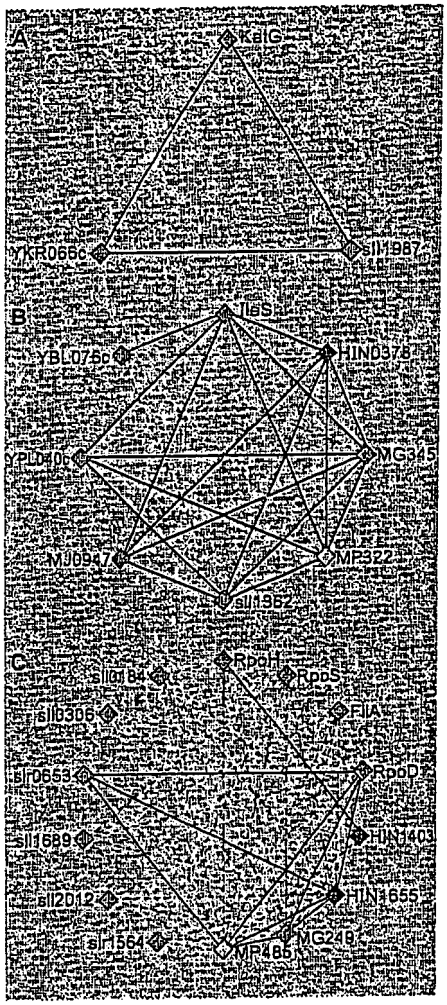


Fig. 1. Examples of COGs. Solid lines show symmetrical BeTs. Broken lines show asymmetrical BeTs, with color corresponding to the species for which the BeT is observed. Genes from the same species are adjacent; otherwise the gene names are positioned arbitrarily. A unique COG ID is indicated in the upper left corner. (**A**) Congruent BeTs form a triangle, the minimal COG. Origin of the proteins: KatG, *E. coli*; sll1987, *Synechocystis* sp.; and YKR066c, *S. cerevisiae*. Note that all the BeTs are symmetrical. (**B**) A simple COG with two yeast paralogs. Origin of the proteins: IleS, *E. coli*; HIN0378, *H. influenzae*; MG345, *M. genitalium*; MP322, *M. pneumoniae*; MJ0947, *M. jannaschii*; and YBL076c and YPL040c, *S. cerevisiae*. Note the adjacent triangles with a common side, for example, IleS-MG345-MJ0947 and sll1362-MG345-MJ1362. YPL040c is the yeast mitochondrial isoleucyl-tRNA synthetase; the bacterial orthologs and that from *M. jannaschii* are the BeTs for this yeast protein, but the reverse is true only of the bacterial proteins (symmetrical BeTs). Conversely, for YBL076c, which is the yeast cytoplasmic isoleucyl-tRNA synthetase, the *M. jannaschii* ortholog is a symmetrical BeT, whereas the bacterial BeTs are asymmetrical. (**C**) A complex COG with multiple paralogs. Origin of the proteins: RpoH, RpoS, RpoD, and FliA, *E. coli*; HIN1403 and HIN1655, *H. influenzae*; MG249, *M. genitalium*; MP485, *M. pneumoniae*; sll0184, sll0306, slr0653, sll1689, sll2012, and slr1564, *Synechocystis* sp. RpoD, HIN1655, slr0653, and MG249 are major sigma factors (σ70), whose function is universal in bacteria; note the fully symmetrical relationships between these proteins. The other proteins are specialized sigma factors whose radiation from the ancestral family apparently was accompanied by modification of the function and involved accelerated evolution; note the asymmetrical BeTs.

COGs whose corresponding paralogs showed consistently lower similarity to other members of the COG. One may think of the rapidly evolving paralogs as progenitors of new families emerging from within the conserved ones. The COGs will be an important resource in a systematic survey of the functional diversification of paralogs in conserved gene families.

There are several large clusters in the current collection with complex relationships between members. Two of these, namely the adenosine triphosphatase (ATPase) components of ABC transporters and histidine kinases, each include over 100 members. It is likely that subsequent detailed analysis of these large groups (for example, by phylogenetic tree methods) will result in their split into several distinct COGs, especially when more genomes are available. On a more general note, COGs do not supplant traditional methods of phylogenetic analysis but rather provide the appropriate starting material for these methods, in particular for a systematic analysis of phylogenetic tree topology.

Figure 2 shows the breakdown of the COGs by broadly defined function (*19*) and by species (*20*). For the majority of the COGs, the protein function is either known from direct experiments, mainly in *E. coli* or yeast, or can be confidently inferred on the basis of significant sequence similarity to functionally characterized proteins from other species. It has to be emphasized that construction of the COGs includes automatic prediction of the function for numerous genes, particularly from the poorly characterized genomes such as *M. jannaschii*. There is, however, a substantial fraction of the COGs (14%) for which only general functional prediction, typically of biochemical activity, but not the actual cellular role could be made, and for another 5%, there was no functional clue (Fig. 3). Each of the COGs includes proteins from at least three major clades whose divergence time is estimated to be over a billion years (*21*), that is, they all are ancient, conserved families with important, if not necessarily essential, cellular functions. Therefore, the proteins belonging to the "mysterious" COGs are good candidates for directed experimental studies.

The distribution of proteins from different species in the COGs shows several trends (Fig. 2), although the bias in the current collection of complete genomes (in particular, because three lineages are required to form a COG, all COGs had to have a bacterial member) must be taken into account when interpreting these comparisons. The fraction of proteins belonging to COGs is greatest in the nearly minimal genomes of mycoplasmas (70% for *M. genitalium*) and much lower in the larger genomes of *E. coli* and yeast (40% and 26%, respectively), which indeed is the tendency expected of conserved families presumably associated with cellular housekeeping functions. The genes of the pathogenic bacteria (*H. influenzae* and two mycoplasmas) are essentially subsets of the two larger bacterial gene complements, *E. coli* and *Synechocystis* sp. The latter two species almost always cooccur in the COGs. The main cause of the observed congruency is likely to be the conservation of the core of ancestral bacterial genes in nonparasitic species from different major clades. Accordingly, the fact that proteins from the pathogenic bacteria are missing in many COGs most likely testifies to gene loss, which has been extensive even in this subset of highly conserved genes. The cooccurrence of *M. jannaschii* in a COG with *E. coli* or *Synechocystis* is measurably more frequent than that with yeast (Fig. 2). Such a distribution of the archaeal genes appears to be due primarily to the blending of bacterial-like and eukaryotic-like genes in the archaeal genomes (*10*), although the mentioned bias in the genome collection is also a factor.

The phylogenetic distribution of the COG members is distinct for different functional classes (Fig. 2). It is not unexpected that translation is the only category in which ubiquitous COGs are predominant. Another obvious trend is the absence of proteins from pathogenic bacteria (*H. influenzae* and, particularly, the mycoplasmas) in many COGs

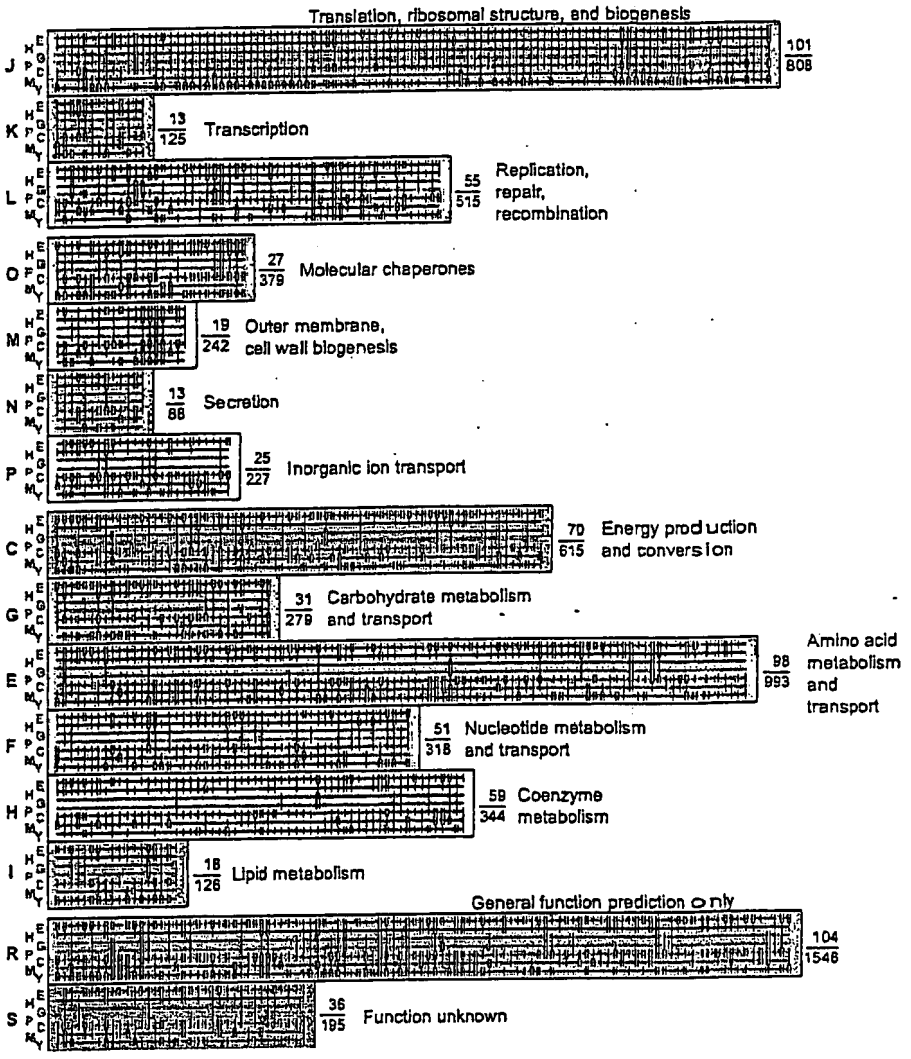


Fig. 2. A functional and phylogenetic breakdown of the COGs. E indicates *E. coli*; H, *H. influenzae*; G, *M. genitalium*; P, *M. pneumoniae*; C, *Synechocystis* sp.; M, *M. jannaschii*; and Y, *S. cerevisiae*. Each column shows a COG; a double streak indicates that two or more paralogs from the given species belong to the particular COG. The number of COGs (numerator) and the number of proteins in them (denominator) is indicated for each functional category. Capital letters in the left-most field encode the functional categories (used in the COG IDs).

in each functional category other than translation and transcription, but especially in the metabolic functional classes. Conversely, the congruence between the two nonparasitic bacteria, *E. coli* and *Synechocystis* sp., holds for all functional classes (Fig. 2). Also apparent is the differential appearance of archaeal proteins that tend to group with yeast proteins in the translation and transcription classes (which, given the bias in the genome collection, results in ubiquitous COGs) but in all other functional classes are frequently found in COGs with bacterial proteins only.

The phylogenetic distribution of COG membership can be conveniently presented in terms of "phylogenetic patterns," which show the presence or absence of each analyzed species (Fig. 3). Of the 88 patterns that include at least three lineages (the definition of a COG), 36 were actually found. Missing were mostly patterns with only one of the two species of *Mycoplasma*, which was predictable because the gene complement of *M. genitalium* is essentially a subset of the *M. pneumoniae* complement (22). The remaining eight patterns that were never observed all include pathogenic bacteria without *E. coli*, which is the largest and most diverse of the available bacterial genomes. The two most abundant patterns could easily be predicted: all species ("ehgpcmy"), and all species except for the mycoplasmas ("eh__cmy"). What appears much less trivial is that these patterns together encompass only one-third of all COGs. This fact emphasizes the remarkable fluidity of genomes in evolution, revealed in spite of the fact that the analysis concentrated on ancient conserved families. Multiple solutions for the same important cellular function appear to be a rule rather than an exception, at least when phylogenetically distant species are considered (*10*, *23*). On the other hand, the eight most frequent patterns, which together account for 85% of the COGs, all include both *E. coli* and *Synechocystis*, emphasizing the congruency between these genomes.

The 114 ubiquitous COGs, most of them including components of the translation and transcription machinery, form the universal core of life. This set is more than twofold down from the bacterial "minimal set" consisting of 256 genes (*23*), but significant further erosion seems unlikely, given the broad spectrum of compared genomes.

The higher order distribution of the COGs by the three domains of life, with only 45% of the COGs including representatives of Bacteria, Archaea, and Eukarya, is another manifestation of the dynamics of gene families in evolution (Fig. 3). The picture is expected to become even more complex, and the fraction of three-domain COGs will probably drop, once archaeal-only, eukaryotic-only, and archaeal-and-eukaryotic COGs emerge with the accumulation of genome sequences.

The unusual, rare patterns are of particular interest, suggesting the possibility of unexpected findings. Each of the COGs with patterns that occur only once in our current collection (Table 1) should correspond to a unique function scattered over disconnected branches of the tree of life. Why such functions are conserved and are presumably important for survival in some but not other lineages is a challenge to be addressed experimentally. The principal evolutionary mechanisms that can be invoked to explain the emergence of these rare patterns are differential gene loss and horizontal transfer of genes. Some of the functions involved, for example, lipoate protein ligase and glycyl-transfer ribonuclease (tRNA) synthetase, appear to be strictly essential, but in different species, they are performed by two distinct sets of orthologs unrelated to one another (*24*). Other functions, for example, thymidine phosphorylase and hexuronate dehydrogenases, may be dispensable under most conditions, and accordingly, differential gene loss is likely; it is remarkable, however, that these functions are preserved in the nearly minimal gene complements of the mycoplasmas. Two of the unique patterns, namely "__gpc_y" and "_hgp__y," might have evolved through horizontal transfer of typical eukaryotic genes into bacterial genomes. The latter pattern is of particular interest as it involves the choline kinase gene common to a number of bacterial pathogens and implicated in pathogenicity (*25*). Two of the COGs with unique patterns, "h__c_y" and "e_gp_my," include highly conserved but uncharacterized proteins whose functions could be predicted only by detailed analysis of conserved protein motifs (Table 1). These examples demonstrate the potential for protein function prediction inherent in the construction of the COGs themselves.

The sampling of genomes we compared is small and biased, and when a more complete set is available, the distribution of COGs by phylogenetic patterns is likely to change significantly; for example, many patterns that are currently rare may become common when larger genomes from the Gram-positive bacterial lineage (such as *Bacillus subtilis*) become available. Nevertheless, we believe that the language of phylogenetic patterns will become even more useful for the description of relationships between multiple genomes.

Bacteria+Eukarya +Archaea		Bacteria+Eukarya		Bacteria+Archaea		Bacteria only	
Pattern	COGs	Pattern	COGs	Pattern	COGs	Pattern	COGs
						e_gpc__	5
e___cmy	37			ehgpcm_	15	eh_pc__	2
eh___my	18	ehgp__y	5	e_gpcm_	4		
____cmy	13	e_gpc_y	2	_h__cm_	3		
e____my	7	e__p__y	1	eh_p_m_	2		
__gpcmy	4	e_gp__y	1	ehgp_m_	2		
_h___my	2	eh_pc_y	1	e_gp_m_	1		
eh_p_my	2	_h__c_y	1				
ehgp_my	2	__gpc_y	1				
e_gpcmy	2	_hgp__y	1				
__gp_my	1						
e_gp_my	1						
eh_pcmy	1						
Sum	323	215		122		60	
COGs(%)	45	30		17		8	

Fig. 3. Phylogenetic patterns in COGs. Letter codes as in Fig. 2 (ignore case); an underline indicates absence of the respective species. Shading indicates the eight most frequent patterns.

Connecting and Expanding the COGs

Ancient families of paralogs that span a broad range of taxa are well known (*26*). Accordingly, a number of COGs are related to each other and can be connected into superfamilies. In order to elucidate the superfamily structure of the COG collection, we used the recently developed PSI-BLAST (position-specific iterative BLAST) program, which combines BLAST search with profile analysis (*27*). Two COGs were considered connected if at least two of the proteins from the first COG hit members of the second COG in the PSI-BLAST search, and vice versa. Clustering by this criterion produced 58 superfamilies including 280 COGs.

Compared to COGs themselves, the superfamilies are a higher level of protein classification. Typically, they include conserved motifs that are determinants of a distinct biochemical activity, which, however, may be required for a variety of cellular functions. For example, the largest superfamily contains 53 COGs with 863 proteins, all of which contain conserved motifs typical of ATPases and GTPases but are involved in a broad range of processes from DNA replication to metabolite transport (*28*).

Superfamilies and their signature motifs

will be useful in classifying proteins that have evolved to an extent that they cannot be assigned to any COG but still retain a conserved motif. We sought to detect such proteins with distant, subtle similarity to COGs that might be encoded in the analyzed genomes. The PSI-BLAST analysis (27) detected "tails" of distantly related proteins (a total of 3686) for 321 COGs, increasing the total number of proteins connected to COGs to 10,332 (58% of the entire protein set from complete genomes).

Because apparent orthologs from at least three major clades were required to form a COG, there are potential new COGs hidden among the results of the comparison of protein sequences from complete genomes (11). Clustering by sequence similarity the proteins not included in COGs (14) resulted in 443 groups with members from two clades. Predictably, the greatest number, 204, were from the cyanobacterial and Gram-negative clades, followed by 67 groups combining yeast and *M. jannaschii*. Many of these groups are likely to become COGs once additional genomes are included in the analysis.

Prediction of Protein Functions with the COG System

The COG system allows automatic functional and phylogenetic annotation of genes and gene sets (29). As in the procedure used for the construction of the COGs, the criterion for adding likely orthologs from other genomes to the COGs is based on the consistency between the observed relationships. A protein is compared to the database of protein sequences from complete genomes (11) and is included in a COG if at least two BeTs fall into it. Given that the COGs were constructed from proteins encoded in complete genomes, it is not a requirement that newly included proteins also originate from a complete genome. Indeed, while the unsequenced portion of a genome may encode proteins with the highest similarity to those included in COGs, the BeTs will not change for the products of already sequenced genes.

As a demonstration of the principle coupled with additional characterization of the COGs themselves, the sequences of proteins with known three-dimensional structures from the PDB database (30) were compared to the protein sequences encoded in complete genomes. The "two BeT" procedure resulted in proteins with known three-dimensional structure being included in 183 COGs, of which one was shown to be a false positive by subsequent alignment analysis. Thus, structural information could be inferred for at least 25% of the COGs. In most cases, the structurally characterized protein (from *E. coli* or yeast) actually belongs to a COG or is a closely related homolog of the proteins forming a COG.

Some of the predictions, however, provide significant functional and structural inferences. Of particular interest are (i) the possibility of modeling the nuclease domain of polyadenylate cleavage factors

Table 1. Unique phylogenetic patterns among COGs. The pattern designations are as in Fig. 3; each COG ID includes a letter indicating the functional category, to which the constituent proteins belong (Fig. 2).

Pattern and COG ID	Proteins	Activity or function	Comment
e_gp_m_ COG0213F	DeoA-MG051-MP090-MJ0667	Thymidine phosphorylase; salvage of deoxypyrimidines	Nonessential gene in *E. coli*; apparent orthologs found in other Gram-positive bacteria and in humans (35).
e__p__y COG0246G	MtlD, UxaB, UxuB, YdfI, YeiQ-MP190-YEL070w, YNR073c	Mannitol-1-phosphate and other hexuronate dehydrogenases; hexuronate catabolism	Nonessential genes in *E. coli*; accessory reactions of carbohydrate metabolism (36).
e_gp__y COG0095H	LplA-MG270-MP450-(sll0809)-YJL046w	Lipoate-protein ligase A; ligation of lipoate to apoproteins of pyruvate dehydrogenase and other lipoate-dependent enzymes	There are two unrelated classes of lipoate-protein ligases; *E. coli* and yeast encode both forms; *H. influenzae* and *Synechocystis* sp. encode the B form (included in a separate COG); sll0809 is a distant homolog of the A form (37), which was not automatically included in the COG but was detected with PSI-BLAST.
eh_pc_y COG0604R	AdhC + 18 *E. coli* proteins–MP278-sll0990, slr1192-YBR046c + 19 yeast proteins	Alcohol dehydrogenase class III and related Fe-S dehydrogenases; various catabolic pathways	Highly conserved protein family distinct from other Fe-S oxidoreductases.
_h__o_y COG0678R	HIN1693_1-sll1621-YLR109w	Glutaredoxin-like membrane protein (prediction)	The *H. influenzae* protein contains an additional thioredoxin-like domain.
__gpc_y COG0631R	MG108-MP586-sll1771-sll1033-sll0602-YDL006w + 6 yeast proteins	Protein serine and threonine phosphatase	Serine and threonine protein phosphatases are abundant in eukaryotes but not in bacteria (38).
__gp_my COG0423J	MG251-MP483-MJ0228-YPR081c, YBR121c	Glycyl-tRNA synthetase (eukaryotic and Gram-positive type)	Gram-negative bacteria and *Synechocystis* encode a distinct glycyl-tRNA that appears to be unrelated to the eukaryotic and Gram-positive type; the closest relative of this COG in *E. coli* and *H. influenzae* is prolyl-tRNA synthetase (24).
e_gp_my COG0622R	b2300-MG207, MP029-MJ0623, MJ0936-YHR012w	Phosphoesterase (prediction)	Highly conserved protein family that shares only modified catalytic motifs (detected by PSI-BLAST; $P \sim 0.004$) with other phosphoesterases, including protein phosphatases.
eh_pcmy COG0078E	ArgI, ArgF, YgeW-HIN0012-MP531-sll0902-MJ0881-YJL088w	Ornithine carbamoyltransferase; arginine biosynthesis	Amino acid metabolism appears to be completely missing in *M. genitalium*, but residual reactions may occur in *M. pneumoniae*.
_hgp_y COG0510M*	HIN0938-MG356, MP310-YDR147w, YLR133w	Choline kinase (prediction) involved in lipopolysaccharide biosynthesis	Enzyme common to several bacterial pathogens and eukaryotes; contributes to pathogenicity (25).

*This COG was added to the collection by cluster analysis.

(31) with the beta-lactamase structure, (ii) the presence of an acylphosphatase domain in hydrogenase expression factors, which form a highly conserved COG, and in a number of uncharacterized proteins, and (iii) the connection between a unique carbonic anhydrase and an acetyltransferase family (Table 2).

Probably the most important application of the COGs is functional characterization of newly sequenced genomes. In the preliminary analysis of the recently published genome of the major human bacterial pathogen *Helicobacter pylori* (32), 813 proteins (51% of the gene products) from this bacterium were included in 453 preexisting COGs and 143 new COGs (33). In spite of the fact that many *H. pylori* proteins are highly similar to homologs from *E. coli* and other bacteria and have been explored in detail (32), this analysis produced over 100 additional functional predictions (33).

Conclusions and Perspective

The COGs bring together the fields of comparative genomics and protein classification. Among the numerous possible approaches to protein classification, the COGs appear to be unique as a prototype of a natural system, which has as its basic unit a group of descendants of a single ancestral gene. Typically, such a group is associated with a conserved, specific function, so that the inclusion of a protein in a COG automatically entails functional prediction.

Each COG contains conserved genes from at least three phylogenetically distant clades and, accordingly, corresponds to an ancient conserved region (ACR). Previous analyses have indicated that the total number of distinct ACRs is likely to be less than 1000 (34). Thus, even with the limited number of complete genomes currently available for analysis, the COGs have already captured a substantial fraction of all existing highly conserved protein domains. With more genomes included in the system, the discovery of additional COGs should gradually level off, with the great majority of the ACRs encoded in the added genomes fitting into already known COGs.

With the forthcoming flood of genome sequences, a coherent framework for understanding these genomes from both the functional and evolutionary viewpoints is a must. We regard the current collection of

Table 2. Structural and functional predictions for uncharacterized proteins in COGs.

Phylogenetic pattern and COG ID*	Proteins in COG†	Activity and function	Homolog in PDB‡ ·BeTs detected (no.) ·Lowest *P* with a COG member	Comment
e_gpcmy COG0595R	PhnP, ElaC-2g-2p-5c-Bm-YLR277c, YMR137c, YKR079c	Predicted Zn-dependent hydrolases	Beta-lactamase (1BMC) ·2 ·0.039	Activity is not known for any protein in this ubiquitous COG. Biochemical and genetic data indicate that YLR277c is involved in messenger RNA 3′-end processing (31), whereas YMR137c is DNA cross-link repair protein SNM1 (39). A motif including the Zn-coordinating histidines of beta-lactamase is conserved.
eh__cmy COG0607R	SseA, PspE, GlpE, YibN, YbbB, YnjE, YgaP-2h-5c-MJ0052-4y	Predicted sulfur-transferases	Rhodanese (1RHD, 2ORA, 1ORB) ·2 ·10^{-41}	The sulfurtransferase activity of SseA has been demonstrated (40), but the rest of the proteins in this COG have no known activity. PspE (phage shock protein), GlpE (uncharacterized protein involved in glycerol metabolism), and other small proteins correspond to one of the two rhodanese domains.
ehgpc_y COG0596R	PldB, MhpC, YcdJ, YnbC-HIN0065-MG020-MP132-6c-YNR064c, YKL094w	Predicted hydrolases and acyltransferases	Lipases (2LIP, 1TAH|B, 1CVL) ·3 ·8×10^{-5}	PldB is known to possess triglyceride lipase activity (41). All other proteins in the COG have not been characterized but now can be predicted to possess the α- or β-hydrolase fold.
e___cm_ COG0068C	HypF-sll0322-MJ0713	Hydrogenase maturation factor	Acylphosphatase (1APS) ·2 ·2×10^{-5}	HypF is required for hydrogenase biosynthesis (42), but no biochemical activity is known. The ~100 amino acid, NH_2-terminal domain aligns with acylphosphatase, with the catalytic residues conserved, suggesting that HypF orthologs indeed possess acylphosphatase activity. A PSI-BLAST search with this domain as the query detected five additional likely acylphosphatases, namely *E. coli* YccX and *M. jannaschii* MJ0809, MJ0553, MJ1331, and MJ1405 (43).
e___cm_ COG0663R	CaiE, YrdA, YdbZ-sll1636, sll1031-MJ0304	Predicted carbonic anhydrases	Carbonic anhydrase from Methanosarcina thermophila (1THJ) ·3 ·10^{-29}	The biochemical activity of the proteins in this COG is not known. They show not only conservation of histidine residue comprising the active center of this unusual carbonic anhydrase (44) but also significant similarity to acetyltransferases of the isoleucine patch superfamily (45), suggesting an unexpected connection between the two types of enzymes.

*The designations are as in Table 1 and Fig. 3. †2g indicates two proteins from *M. genitalium*, 2p indicates two proteins from *M. pneumoniae*, and so forth. ‡The PDB accession is indicated in parentheses.

COGs as a crude first version of such a framework. Inclusion of additional, phylogenetically diverse genomes and further development of the procedures used to derive and analyze COGs will hopefully result in refinement of this system, making it a solid platform for genome annotation and evolutionary genomics.

REFERENCES AND NOTES

1. R. D. Fleischmann *et al.*, *Science* 269, 496 (1995).
2. C. M. Fraser *et al.*, *ibid.* 270, 397 (1995); R. Himmelreich *et al.*, *Nucleic Acids Res.* 24, 4420 (1996); T. Kaneko *et al.*, *DNA Res.* 3, 109 (1996); F. R. Blattner *et al.*, *Science* 277, 1453 (1997).
3. C. J. Bult *et al.*, *Science* 273, 1058 (1996).
4. A. Goffeau *et al.*, *ibid.* 274, 546 (1996); H. W. Mewes *et al.*, *Nature* 387, 7 (1997).
5. C. R. Woese, *Curr. Biol.* 6, 1060 (1996); G. J. Olsen and C. R. Woese, *Cell* 89, 991 (1997); E. V. Koonin, *Genome Res.* 7, 418 (1997).
6. E. V. Koonin, A. R. Mushegian, K. E. Rudd, *Curr. Biol.* 6, 404 (1996); E. V. Koonin and A. R. Mushegian, *Curr. Opin. Genet. Dev.* 6, 757 (1996).
7. W. M. Fitch, *Syst. Zool.* 19, 99 (1970). This definition may not embrace all of the complexity of relationships between genes in different genomes. For example, if genes A and B are paralogs encoded in genome 1, and A′ and B′ are their respective orthologs in genome 2, what is the appropriate description of the relationship between A and B′? They formally are not paralogs, even though a generalized definition might include such cases. Furthermore, one-to-many and many-to-many orthologous relationships evidently exist.
8. W. M. Fitch, *Philos. Trans. R. Soc. London Ser. B* 349, 93 (1995).
9. R. L. Tatusov *et al.*, *Curr. Biol.* 6, 279 (1996).
10. E. V. Koonin, A. R. Mushegian, M. Y. Galperin, D. R. Walker, *Mol. Microbiol.* 25, 619 (1997).
11. The protein sequences were from the original references (*1*–*4*), with modifications (for example, tentative correction of frame-shift errors) and additions (previously unreported predicted genes) made for *E. coli* (E. V. Koonin and R. L. Tatusov, unpublished observations; K. E. Rudd, personal communication), *H. influenzae* (*9*), *M. genitalium* and *M. jannaschii* (*10*), and *S. cerevisiae* (T. J. Wolfsberg and D. Landsman, personal communication). The list of systematic names for all *E. coli* genes was provided by K. Rudd, and the names for all yeast genes were provided by T. Wolfsberg and D. Landsman; the *H. influenzae* genes were renamed as previously described (*9*); the gene names for the other species were from the original publications. The resulting protein database from complete genomes used in all comparisons contained 4283 sequences from *E. coli*, 1703 sequences from *H. influenzae*, 468 sequences from *M. genitalium*, 677 sequences from *M. pneumoniae*, 3168 sequences from *Synechocystis* sp., 1736 sequences from *M. jannaschii*, and 5932 sequences from *S. cerevisiae*, totaling 17,967 sequences. This sequence set is available on the World Wide Web at http://www.ncbi.nlm.nih.gov/COG. All pairwise comparisons between these sequences were performed using the BLASTPGP program, which is based on an enhanced version of the BLAST algorithm and includes analysis of local alignments with gaps (*26*). Predicted coiled coil regions in protein sequences were masked before the comparison using the batch version of the COILS2 program [A. Lupas, *Methods Enzymol.* 266, 513 (1996); D. R. Walker and E. V. Koonin, *ISMB* 5, 333 (1997)], and additionally, regions of low complexity were masked using the SEG program with default parameters [J. C. Wootton and S. Federhen, *Methods Enzymol.* 266, 554 (1996)]. Before the detection of triangles of BeTs, paralogs were identified as those proteins from the same lineage that showed greater similarity to each other than to any protein from another lineage. For the purpose of triangle formation, paralogs were treated as a group. The algorithm further included verification that the BeTs included in a triangle formed a consistent multiple alignment; triangles that did not contain a conserved motif were disregarded.
12. Although the exact solution depends on the amino acid composition and size of the particular proteins, under zero approximation, if B (from genome *b*) is the BeT for A (from genome *a*), and C (from genome *c*) is the BeT for B, the probability that C is the BeT for A by chance is close to $1/N$, where N is the number of genes in genome *c*, or ~0.001.
13. C. R. Woese, *Microbiol. Rev.* 51, 221 (1987); ———, R. Overbeek, G. J. Olsen, *J. Bacteriol.* 176, 1 (1994); N. R. Pace, *Science* 276, 734 (1997). A BeT to a given clade was registered if detected in any of the constituent species, for example, in *E. coli* or *H. influenzae* for the Gram-negative bacteria.
14. H. Watanabe and J. Otsuka, *Comput. Appl. Biosci.* 11, 159 (1995); E. V. Koonin, R. L. Tatusov, K. E. Rudd, *Methods Enzymol.* 266, 295 (1996).
15. A schematic visual representation of the search results was used for this analysis [T. L. Madden, R. L. Tatusov, J. Zhang, *Methods Enzymol.* 266, 131 (1996)].
16. A single-linkage clustering procedure was used with random match probability, $P < 0.001$, as the cutoff (*14*).
17. A searchable database of COGs is available at http://www.ncbi.nlm.nih.gov/COG. Each COG was assigned a unique identification number, which includes a letter for the functional category (*19*) and a number (see examples in Fig. 1 and Tables 1 and 2).
18. M. Lonetto, M. Gribskov, C. A. Gross, *J. Bacteriol.* 174, 3843 (1992).
19. The broad functional categories of proteins were as defined previously (*9*), except that transcription was separated from replication, recombination, and repair. This classification is a modification of the system originally developed for *E. coli* proteins [M. Riley, *Microbiol. Rev.* 57, 862 (1993)].
20. A partially similar representation of some of the protein families from complete genomes has been recently published [R. A. Clayton, O. White, K. A. Ketchum, J. C. Venter, *Nature* 387, 459 (1997)].
21. R. F. Doolittle, D.-F. Feng, S. Tsang, G. Chao, E. Little, *Science* 271, 470 (1996).
22. R. Himmelreich *et al.*, *Nucleic Acids Res.* 25, 701 (1997).
23. A. R. Mushegian and E. V. Koonin, *Proc. Natl. Acad. Sci. U.S.A.* 93, 10268 (1996).
24. E. V. Koonin, A. R. Mushegian, P. Bork, *Trends Genet.* 12, 334 (1996).
25. J. N. Weiser, M. Shchepetov, S. T. Chong, *Infect. Immun.* 65, 943 (1997).
26. J. P. Gogarten *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 86, 6661 (1989); N. Iwabe *et al.*, *ibid.*, p. 9355; J. P. Gogarten, E. Hilario, L. Olendzewski, in *Evolution of Microbial Life*, D. McL. Roberts, P. Sharp, G. Alderson, M. Collins, Eds. (Cambridge Univ. Press, Cambridge, 1996), pp. 267–292.
27. S. F. Altschul *et al.*, *Nucleic Acids Res.* 25, 3389 (1997). The probability of a random match, $P < 0.001$, was used in all PSI-BLAST searches.
28. J. E. Walker, M. Saraste, M. J. Runswick, N. J. Gay, *EMBO J.* 1, 945 (1982); A. E. Gorbalenya and E. V. Koonin, *Nucleic Acids Res.* 17, 8413 (1989); M. Saraste, P. R. Sibbald, A. Wittinghofer, *Trends Biochem. Sci.* 15, 430 (1990).
29. Protein sequences can be submitted for searching against COGs at http://www.ncbi.nlm.nih.gov/COG/cognitor.html
30. F. C. Bernstein *et al.*, *J. Mol. Biol.* 112, 535 (1977).
31. G. Chanfreau, S. M. Noble, C. Guthrie, *Science* 274, 1511 (1996); A. Jenny, L. Minvielle-Sebastia, P. J. Preker, W. Keller, *Ibid.* 274, 1514 (1996); G. Stumpf and H. Domdey, *Ibid.*, p. 1517.
32. J.-F. Tomb *et al.*, *Nature* 388, 539 (1997).
33. E. V. Koonin, R. L. Tatusov, M. Y. Galperin, M. N. Rozanov, unpublished observations.
34. P. Green *et al.*, *Science* 259, 1711 (1993).
35. J. Neuhard and R. A. Kelln, in Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, ed. 2, 1996), pp. 580–599.
36. E. C. C. Lin, *ibid.*, pp. 307–342.
37. T. W. Morris, K. E. Reed, J. E. Cronan Jr., *J. Bacteriol.* 177, 1 (1995).
38. P. Bork, N. P. Brown, H. Hegyi, J. Schultz, *Protein Sci.* 5, 1421 (1996).
39. D. Richter, E. Niegemann, M. Brendel, *Mol. Gen. Genet.* 231, 194 (1992); R. Wolter, W. Slede, M. Brendel, *Ibid.* 250, 162 (1996).
40. H. Hama, T. Kayahara, W. Ogawa, M. Tsuda, T. Tsuchiya, *J. Biochem.* 115, 1135 (1994).
41. T. Kobayashi *et al.*, *Ibid.* 98, 101 (1985).
42. A. Colbeau *et al.*, *Mol. Microbiol.* 8,15 (1993).
43. M. N. Rozanov and E. V. Koonin, unpublished observations.
44. B. E. Alber and J. G. Ferry, *Proc. Natl. Acad. Sci. U.S.A.* 91, 6909 (1994); C. Kisker *et al.*, *EMBO J.* 15, 2323 (1996).
45. E. V. Koonin, *Protein Sci.* 4, 1608 (1995); M. N. Rozanov and E. V. Koonin, unpublished observations.

46. We thank A. Schaffer for modifying the PSI-BLAST program; R. Walker, H. Watanabe, and M. Rozanov for valuable help with data analysis; K. Rudd, T. Wolfsberg, and D. Landsman for unpublished data; and P. Bork, M. Galperin, M. Gelfand, A. Mushegian, P. Pevzner, M. Roytberg, M. Rozanov, and R. Walker for helpful discussions.

Microbial pathogen genomes – new strategies for identifying therapeutics and vaccine targets

Douglas R. Smith

Advances in high-throughput DNA-sequencing techniques have given us the unprecedented ability to rapidly determine the nucleotide sequences of entire bacterial genomes. The application of these methods to the genomes of microbial pathogens, combined with efficient analytical tools and genome-scale approaches for studying gene expression, is revolutionizing our approach to the selection of targets for drug screening and vaccine development. This is bringing new life to this important, but long-neglected, field of research.

The decision, several years ago, by the US Department of Energy, the National Institutes of Health (NIH) and several international funding agencies to embark upon programs to map and sequence the human genome has led to a number of important technological advances that are beginning to have an impact in other areas of biology. Among these advances are the development of automated methods for the generation of large amounts of raw DNA-sequencing information, computer software for rapidly processing and analyzing primary sequence data, and techniques for the rapid assembly of shotgun sequencing reads, even from entire bacterial genomes. Efficient algorithms for similarity searching allow the rapid identification of protein-encoding sequences that are homologous to other genes, the sequences of which are held in public and private databases; as from April 1996, approximately 500 megabases (Mb) of nucleotide sequence were contained in GenBank, and approximately 200 000 sequences were held in the SWISS-PROT/Genpept/PIR database of non-redundant proteins. Combined with the wealth of biochemical information that is archived in public databases, it has become possible to describe rapidly the full repertoire of genes in a microbial genome, and to predict many of the metabolic pathways that an organism may utilize.

Progress in this field has been stimulated by the interests of the biotechnology and pharmaceutical industries in using genome-sequencing data as a basis for drug discovery. In turn, this has led to the development of proprietary databases containing genomic information, which provide the basis for *in silico* experiments to identify novel targets for drugs, and for laboratory experiments to identify genes that perform critical functions. This article summarizes some recent developments in this important area, focusing on bacterial sequences, and provides examples to illustrate how genome-sequencing information from microbial pathogens can be used to select targets for vaccine and drug development. The overall process used to proceed from sequence generation to target validation is illustrated in Fig. 1.

D. R. Smith (smith@tric.com) is at Genome Therapeutics Corporation, 100 Beaver Street, Waltham, MA 02154, USA.

Large-scale sequencing of bacterial genomes

Many laboratories use automated sample-preparation techniques and fluorescence-based gel readers [such as that produced by Applied Biosystems Inc., (ABI); Foster City, CA, USA] for the large-scale sequencing of bacterial genomes. These instruments have the advantage that they are efficient, and relatively easy to set up and operate. A few laboratories use computer-assisted multiplex sequencing to achieve the same end[1]. In multiplex sequencing, samples consisting of pools of up to 20 plasmids are processed through sample preparation and gel electrophoresis, and the resulting sequences are determined from electroblots of the gels by hybridization with radioactive or fluorescently labeled probes. This technique can be used to generate 40 films (or digitized images) from each sequencing gel. Although multiplex sequencing is efficient at producing large amounts of 'shotgun' data, it is more difficult to set up and operate in the laboratory than is fluorescence-based gel sequencing, and it is not suited to directed-finishing strategies. ABI machines are used in the author's laboratory to generate primer-directed reads for finishing and gap closure.

During the past year, a group at The Institute for Genomic Research (TIGR; Gaithersburg, MD, USA) reported the complete sequences of *Haemophilus*

influenzae (1.8 Mb), a major cause of respiratory infections and meningitis, especially in children[2], and of *Mycoplasma genitalium* (0.6 Mb), which causes urethritis[3]. Approximately 1.6 Mb of contiguous sequence from the 4.7 Mb *Escherichia coli* genome has been published[4], and the sequencing of a further 2 Mb was reported at the 1995 Genome Sequencing and Analysis VII (GSA-VII) meeting[5]. The genome of *Helicobacter pylori* (1.7 Mb), the major cause of stomach ulcers, has been sequenced by Genome Therapeutics Corporation (GTC; Waltham, MA, USA) under a privately funded microbial-pathogen sequencing program. More than half (1.5 Mb) of the 2.8 Mb genome of *Mycobacterium leprae* (the etiologic agent of leprosy) has also been sequenced by GTC, and is available through GenBank, the GTC web site <http//:www.cric.com>, and through MycDB <http://www.biochem.kth. se/MycDB.html>, which contains mycobacterial genome mapping and sequence information[6].

Other microbial pathogens that are currently being sequenced include *Neisseria gonorrhoeae* (University of Oklahoma, Norman, OK, USA), *Streptococcus pyogenes* (University of Oklahoma), *Treponema pallidum* (University of Texas, Houston, TX, USA, and TIGR), *Mycobacterium tuberculosis* (GTC and the Sanger Centre, Hinxton, Cambridge, UK), and *Staphylococcus aureus* [GTC, and Human Genome Sciences (HGS; Rockville, MD, USA)].

In addition to these pathogens, the genomes of several archaebacteria and other non-pathogens are being sequenced. These include *Methanococcus janaschii* (TIGR), *Pyrococcus furiosis* (University of Utah, Salt Lake City, UT, USA), *Sulfolobus solfataricus* (Dalhousie University, Halifax, Nova Scotia, Canada), and *Pyrobaculum aerophilum* (California Institute of Technology, Pasadena, CA, USA, and University of California, Los Angeles, CA, USA). The 1.7 Mb genome of the archaeon *Methanobacterium thermoautotrophicum* is near completion at GTC (Ref. 7). Approximately 2 Mb of the 4.1 Mb *Bacillus subtilis* genome has now been sequenced by a consortium of European and Japanese laboratories, and the project may be completed by the end of 1996 (Ref. 8). Approximately 1 Mb of genomic sequence from the 2.7 Mb genome of the cyanobacterium *Synechocystis* sp. 6803 was recently published[9].

Within the next couple of years, therefore, we can expect an explosion of bacterial-genome sequence information from species representing a variety of phylogenetic lineages, including many pathogens.

Pharmaceutical companies have shown considerable interest in using pathogen genomics to facilitate the development of vaccines and small-molecule therapeutics. For example, researchers at GlaxoWellcome have sequenced a substantial fraction of the *H. pylori* genome to assist in the process of drug discovery. Over the past year, GTC has formed two research alliances with pharmaceutical companies to take advantage of sequences from microbial pathogens: one with

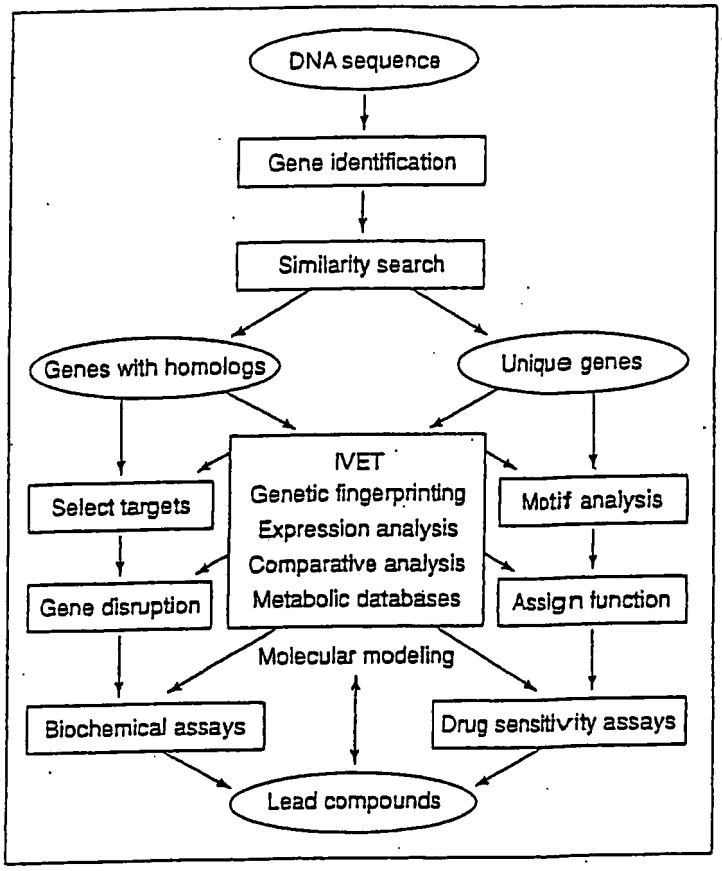


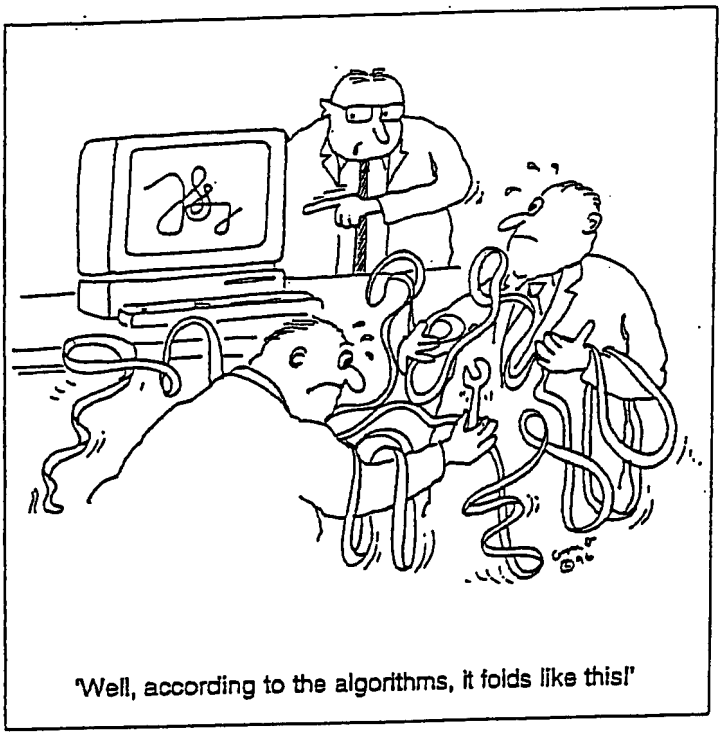
Figure 1

Flow diagram illustrating the process by which a microbial genome sequence is analysed and the information is used to direct experiments and aid in target selection for therapeutics development. The individual steps are referred to throughout the text. In the case of vaccine candidates, gene products from selected targets are expressed and tested in animal models.

biotics and vaccines to treat *H. pylori* infection, and one with Schering-Plough (Union, NJ, USA), to develop broad-spectrum antibiotics and vaccines. Although the genomic route to drug discovery for bacterial pathogens is new and remains unproved, the basic paradigm (outlined below) of gene identification, followed by functional analysis and drug screening, is well established. Thus, it is likely that more companies will become involved, and that in the the future, additional research alliances between genomics companies and the pharmaceutical industry will materialize in this area.

From sequence to genes

The first task when confronted with an entire bacterial-genome sequence, is to identify all the genes. This can be accomplished using a variety of techniques, but the most successful approaches use a combination of reading-frame and codon-usage analysis, together with similarity searching, to identify putative genes with homology to previously described sequences. Commonly used tools include GeneMark[10], GenomeBrowser[11], BLAST (Ref. 12), and highly



'Well, according to the algorithms, it folds like this!'

alignment, such as BLAZE, or MPsrch (Ref. 13). In general, organism-specific codon usage is highly predictive for bacterial genes, but its effective use depends on the existence of sufficient information to generate accurate codon-usage matrices. In some cases, subsets of genes within an organism will exhibit codon-usage patterns that deviate significantly from the norm[14]. Such genes are thought to represent evolutionarily recent acquisitions by phage transduction, conjugation, or some other form of horizontal transfer from other organisms. If enough of these genes are present, codon-usage tables of genomic subsets can be constructed to identify them. Translational start sites can be identified by the occurrence of start codons that coincide with abrupt changes in codon usage, the initiation of homology to previously characterized genes, or the presence of Shine–Dalgarno sequences[13]. Automated analysis tools (such as GenomeBrowser[11]) that provide a graphical display of open reading frames (ORFs), codon usage, database homologies and other features, make the task of identifying bacterial genes and their relationships with each other in the genome relatively straightforward. With the increasing pace of bacterial-genome sequencing, there is an emerging need for second-generation tools that will automate most of the laborious annotation process.

From genes to function

The second phase in the analysis of bacterial genomes is to identify the function of as many genes as possible. Currently, sequence homology is the most powerful tool. A high degree of homology between the putative translation product of a newly identified gene and an enzyme whose function has been thoroughly studied in other organisms, provides strong support for the function of that protein, especially if it is the only homolog in the genome under scrutiny. Other useful tools include programs that identify sequence motifs from databases such as PROSITE (Ref. 16), BLOCKS (Ref. 17), BEAUTY (Ref. 18) and ProDom (Ref. 19). If one is attempting to identify vaccine candidates, then examining highly expressed cell-surface proteins is relevant, so it is then useful to know whether a protein contains a secretion signal, even if nothing else is known about it. Although the tools described here are very good at identifying homologies, 25–40% of the genes in a bacterial genome typically fail to show significant similarity with known proteins.

Once the set of similarity-searching tools has been exhausted, one must return to molecular biology to further elucidate the function and expression pattern of predicted genes. Commonly used approaches to identifying essential genes in an organism include: the use of gene knockouts, disruptions using transposon-mediated mutagenesis, or homologous recombination with disrupted gene-constructs that contain an antibiotic-resistance cassette. Gene disruptions can be generated in a variety of ways, including sophisticated 'hit-and-run' approaches that interrupt a gene without introducing polar effects into downstream ORFs (Ref. 20). However, a gene-by-gene approach to the study of a whole genome is certainly time consuming and labor intensive.

The availability of large amounts of genome-sequence information has stimulated the development of new approaches to functional analysis on a genomic scale. This has been particularly true for researchers investigating yeast, where a concerted effort is being made to ascertain the function of every ORF in the genome. Such strategies include the conceptually simple, but technologically advanced, technique of making microarrays of polymerase chain reaction (PCR)-amplified gene sequences on glass slides to allow the fluorescence-based detection of quantitative hybridization signals from labeled cDNA probes on large numbers of genes simultaneously — perhaps even all the genes of an organism[21]. An ingenious PCR-based approach to efficient sequence-signature-based expression analysis has recently been demonstrated[22]. For example, a technique termed 'genetic fingerprinting' promises to replace individual gene knockouts by a global transposon-mutagenesis approach[23]. Insertions are induced *en masse* in a strain of interest, the strain is grown under a variety of conditions, and PCR products are analysed to identify genes in which transposon hops are under-represented because the genes are required for growth[23]. A conceptually similar dropout technique, which uses tagged transposons to identify the *Salmonella typhimurium* genes required for virulence in a mouse model, has been described[24].

Techniques that probe subsets of genes for a specific functionality, such as secretion or induction during growth in the host, have also been described. These techniques provide clones from which signature

sequences can be derived, so that corresponding genes can be identified by comparing them with the genomic sequence. The IVET (*in vivo* expression technology) technique, which detects gene fusions that result in the *in vivo* selectable expression of a defective *purA* gene or antibiotic-resistance marker, has been used to identify *Salmonella* genes, the expression of which is induced when the pathogen is grown in mice[25]. Finally, protein microsequencing[26] and mass-spectrometry-based peptide analysis[27] have been used to identify protein components (e.g. outer-membrane proteins) in partially purified mixtures, or to identify specific proteins separated by two-dimensional gel electrophoresis. Sequences generated in this manner can be used to correlate specific proteins with the gene sequences from which they are expressed.

Target selection and validation

The techniques described in the previous section can be used to identify genes in specific functional categories that may represent good targets for drug or vaccine development. In general, when developing new antibiotics, one is interested in genes that are essential under all growth conditions (and preferably even in quiescent cells), and for which inhibitors with useful chemical properties, such as permeability and low toxicity, can be identified. One advantage of having the entire sequence of a genome is that targets can be prioritized in terms of their activities and the properties of compounds that are known to interact with them. Even with the results of knockout or *in vivo* expression experiments, additional biological information can aid in narrowing down the field of choices. For example, genes can be selected on the basis of their probable roles in intracellular metabolism. Databases, such as EcoCyc (Ref. 28) or PUMA (Ref. 29), that describe known metabolic pathways can be helpful in this regard. Detailed structural information about homologs of identified genes (determined using the Protein DataBank[30]) can be used to assist in the molecular modeling of inhibitors (some resources for molecular modeling can be found at Ref. 31).

As more genomes are sequenced, it will become possible to identify genes that are unique to a particular organism or group of organisms, or genes that are conserved in certain groups. Thus, for example, it will be possible to use electronic comparison to identify genes that are present in *H. pylori* but not in other gut-dwelling bacteria such as *E. coli*, providing a basis for the development of antibiotics specific to *H. pylori*. Although combinatorial chemistries promise to speed up our ability to synthesize and screen large numbers of unique chemical entities, the sequence-based approach described here provides an avenue for the rational identification and selection of key targets for therapeutics development. Ultimate validation of the targets will, of course, require additional experiments such as protein expression, biochemical-assay development and animal studies to identify those with the most useful properties or inhibitors.

Acknowledgements

The sequencing of *Mycobacterium leprae* and *M. tuberculosis*, and technology development for multiplex sequencing is supported by a NIH Genome Science and Technology Center grant 1P01-HG1106-01 from the National Center for Human Genome Research. The sequencing of *Methanobacterium thermoautotrophicum* is supported under the Microbial Genome Program by Grant No. DE-FC02-95ER61967 from the Office of Health and Environmental Research of the US Department of Energy. The sequencing of *Helicobacter pylori* and *Staphylococcus aureus* is supported by Genome Therapeutics Corporation. Thanks to Brad Guild for comments on the manuscript.

References

1 Church, G. M. and Kieffer-Higgins, S. (1988) *Science* 240, 185–188
2 Fleischmann, R. D. *et al.* (1995) *Science* 269, 496–512
3 Fraser, C. D. *et al.* (1995) *Science* 270, 397–403
4 Burland, V., Plunkett, G., III, Sofia, H. J., Daniels, D. L. and Blattner, F. R. (1995) *Nucleic Acids Res.* 23, 2105–2119
5 Burland, V., Plunkett, G., III and Blattner, F. R. (1995) in *Genome Science and Technology* 1, P-16, Mary Ann Liebert
6 Bergh, S. and Cole, S. T. (1994) *Mol. Microbiol.* 12, 517–534
7 Smith, D. R. *et al.* (1995) in *Genome Science and Technology* 1, P-48, Mary Ann Liebert
8 Devine, K. (1995) *Trends Biotechnol.* 13, 210–216
9 Kaneko, T. *et al.* (1995) *DNA Res.* 2, 153–166
10 Borodovsky, M. and McInnich, J. (1993) *Comput. Chem.* 17, 123–133
11 Robison, K. R. and Church, G. M. (1995) <http://www.belmont.com/gb.html>
12 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410
13 MPsrch <http://www.ebi.ac.uk/searches/blitz.html>
14 Medigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991) *J. Mol. Biol.* 222, 851–856
15 Shine, J. and Dalgarno, L. (1975) *Eur. J. Biochem.* 57, 221–230
16 Bairoch, A. (1991) *Nucleic Acids Res.* 19, 2241–2245
17 Henikoff, S. and Henikoff, J. G. (1991) *Nucleic Acids Res.* 19, 6565–6572
18 Wordey, K. C., Wiese, B. A. and Smith, R. F. (1995) *Genome Res.* 5, 173–184
19 Sonnhammer, E. L. and Kahn, D. (1994) *Protein Sci.* 3, 482–492
20 Link, A. J. and Church, G. M. <http://twod.med.harvard.edu/labgc/pKO3.html>
21 Schena, M., Shalon, D., Davis, R. D. and Brown, P. O. (1995) *Science* 270, 467–470
22 Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) *Science* 270, 484–487
23 Smith, V., Botstein, D. and Brown, P. O. (1995) *Proc. Natl Acad. Sci. USA* 92, 6479–6483
24 Hensel, M. *et al.* (1995) *Science* 269, 400–403
25 Mahan, M. J. *et al.* (1995) *Proc. Natl Acad. Sci. USA* 92, 669–673
26 Tempst, P., Link, A. J., Riviere, L. R., Fleming, M. and Elicone, C. (1990) *Electrophoresis* 11, 537–553
27 James, P., Quadroni, M., Carafoli, E. and Gonnet, G. (1993) *Biochem. Biophys. Res. Commun.* 195, 58–64
28 Karp, P. D. (1992) *CABIOS* 8, 347–357
29 Gaasterland, T., Maltsev, N., Overbeek, R., Selkov, E. <http://www.mcs.anl.gov/home/compbio/PUMA>
30 Protein DataBank <http://www.pbd.bnl.gov>
31 <http://www.pharmacy.wisc.edu>